Applications of Large Language Models in Mental Health Services: Capabilities, Challenges, and Future Directions

SurveyForge

Abstract— The incorporation of large language models (LLMs) into mental health services represents a significant advancement in addressing global mental health challenges by improving diagnostics, therapeutic interventions, and accessibility. This survey explores various applications of LLMs, emphasizing their ability to analyze unstructured data for early disorder detection, understand emotional trends, and enhance diagnostic accuracy. Key findings include LLMs' proficiency in emulating therapeutic dialogue and facilitating virtual counseling. However, challenges such as biases, generalization issues, and privacy concerns persist. The paper identifies the need for real-world validation, emphasizing multimodal data integration and culturally sensitive designs to address these challenges effectively. Future directions include enhancing multilingual and cross-cultural capabilities and developing robust, explainable AI frameworks to build user trust and align LLM outputs with mental health care standards. Through interdisciplinary collaboration, LLMs can be further optimized to ensure inclusive, ethical, and impactful mental health support in diverse settings, ultimately transforming access and delivery within the field.

Index Terms—multimodal data integration, cross-cultural adaptability, explainable AI frameworks

1 Introduction

Global mental health challenges are intensifying, marked by high prevalence rates of conditions like depression, anxiety, and post-traumatic stress disorder (PTSD), affecting hundreds of millions worldwide. Yet, gaps in accessibility, stigma, and the shortage of mental health professionals impede timely and adequate care delivery. As a scalable and innovative technological approach, artificial intelligence (AI) has increasingly been explored to alleviate these barriers. Within AI's scope, large language models (LLMs) have shown transformative potential in mental health services through their ability to process, understand, and generate human-like text at unprecedented scales. In this context, these models hold promise for enhancing mental health diagnostics, interventions, and support systems, but equally present ethical and technical challenges that require rigorous interrogation.

In the past decade, the convergence of natural language processing (NLP) technologies with healthcare has reshaped how mental health is conceptualized and operationalized. Historical progressions from early rule-based NLP systems to present-day LLMs such as GPT-4 and LLaMA-2 highlight leaps in contextual understanding, sentiment analysis, and linguistic generalization. Early approaches were predominantly limited by domain-specific rule sets or reliance on structured inputs, which constrained their generalizability to unstructured mental health data [1]. The advancements in pretrained frameworks (e.g., BERT and RoBERTa) provided the foundation, but limitations in capturing complex psychological constructs and multimodal data limited their usability in applied mental health care scenarios [2], [3].

Today, LLMs trained on diverse and expansive datasets are emerging as powerful tools for addressing key challenges in mental health care. Their ability to analyze unstructured text from social media, electronic health records (EHRs), and conversational platforms enables early detection of mental illnesses, such as depression and PTSD, often outperforming traditional clinical practices. For instance, computational models leveraging Twitter data have demonstrated the potential to predict the onset of depression months before clinical diagnosis by analyzing linguistic features and temporal patterns [4]. Similarly, domain-specific innovations, such as MentalBERT and ClinicalGPT, have shown improved diagnostic precision when applied to mental disorder detection benchmarks [2], [5]. However, the trade-off between generalization and domain-specific specialization remains critical. Generic LLMs often deliver suboptimal results in clinical contexts due to the lack of alignment with mental health-specific linguistic and cultural nuances [6], [7].

Beyond diagnosis, the applications of LLMs in virtual counseling and cognitive interventions reveal their growing role in therapeutic contexts. Studies acknowledge that LLM-based conversational agents show promise in delivering structured cognitive-behavioral therapy (CBT) techniques and motivational interviewing for stress management and mood disorders [8], [9]. Nevertheless, these agents face limitations in generating deep and authentic empathy, an aspect critical to mental health support. Evaluations reveal how models such as ChatGPT reflect behaviors characteristic of low-quality human therapy, underscoring the necessity for fine-tuned ethical, therapeutic alignment [9]. While tools like Psy-LLM aspire to scale mental health services globally by providing immediate, culturally sensitive support, their dependence on high-quality training data introduces risks of systemic bias and the perpetuation of inequalities [10].

Despite the potential utility of LLMs in broadening

access to mental health care, critical challenges persist in their implementation. Ethical risks, such as preserving user privacy, ensuring data security, and addressing LLM hallucinations, remain unresolved [11]. Additionally, disparities in datasets exacerbate biases along socio-cultural dimensions, with underrepresented groups often receiving less accurate or empathetic responses [12]. Future research must prioritize interdisciplinary collaboration across AI ethics, psychiatry, and social sciences to refine LLMs for equitable and responsible deployment. Complementary methodologies like federated learning and multimodal data integration (e.g., EEG and facial cues) could further enhance the granularity and robustness of mental health assessments [13].

In conclusion, as LLMs ascend as transformative tools in mental health care, their integration must be navigated thoughtfully, balancing their vast potential with pragmatic safeguards. Through targeted specialization, multimodal enhancements, and rigorous ethical oversight, LLMs can emerge as key allies in addressing the global mental health crisis, ensuring an inclusive, equitable, and sustainable impact.

2 Early Detection and Diagnostics in Mental Health

2.1 Leveraging Linguistic Markers for Mental Health Assessment

The identification of linguistic markers as indicators of mental health conditions represents a pivotal breakthrough in computational psychiatry. Large language models (LLMs), leveraging their advanced natural language understanding capabilities, have emerged as powerful tools for analyzing textual and conversational data to predict psychological states. Through detailed examination of linguistic markers—ranging from syntactic structures to semantic content—LLMs offer the potential for early, scalable, and non-invasive mental health assessments.

Linguistic markers include features such as word frequency, sentence complexity, sentiment trajectories, and thematic focus, which collectively provide insights into an individual's mental state. Research has demonstrated that individuals with conditions such as depression or anxiety often show reduced lexical diversity, increased use of negative affective language, and a heightened focus on self-referential terms, such as "I" or "my," indicative of inward-focused attention [14]. LLMs excel at detecting these patterns because they are pre-trained on massive text corpora, allowing them to capture subtle relationships between linguistic features and mental states. For instance, models like GPT and BERT variants have outperformed traditional machine learning techniques in mental health classification tasks by identifying fine-grained nuances in text, such as shifts in sentiment across sentences or contextual variations in emotional tone [15], [16].

Emotion and sentiment analysis play a critical role in the detection of mental health conditions. By analyzing polarity, emotional intensity, and sentiment shifts, LLMs can infer mental health-related tendencies. Studies such as those using MentalRoBERTa highlight the promising use of domain-specific pre-trained models for capturing complex emotional states from written text, identifying markers of sadness, fear, or anger often associated with depressive or anxiety disorders [2]. Further, the integration of these models into digital platforms has shown promise for identifying variations in emotion stability, which correlates with disorders such as bipolar disorder or post-traumatic stress disorder (PTSD) [16].

Temporal linguistic patterns, another key dimension, provide valuable indicators of an individual's psychological trajectory over time. LLM-based temporal analysis tools can monitor progressive changes, such as escalating negativity in tone, reduced usage of complex syntax, or withdrawal from emotionally charged language. These patterns may signal deteriorating mental health or impending crises, such as suicidal ideation [4]. Models like Mental-Alpaca have been fine-tuned with such dynamic features, enabling more precise long-term predictions by correlating temporal text trends with mental health conditions [5]. Crucially, generating insights over repeated temporal snapshots allows clinical tools to function not just reactively but also proactively for early intervention.

However, several challenges persist. False positives, often arising from the inherent ambiguity of language, can lead to overdiagnosis or unnecessary interventions [15]. LLMs also struggle with biases in identifying linguistic markers across diverse cultural and linguistic groups, as training datasets often inadequately represent underprivileged or non-Western populations [14]. Additionally, fine-tuning LLMs to distinguish between genuine mental health signals and situational expressions of temporary distress remains a technical hurdle [12]. Until addressed, these limitations could potentially exacerbate inequalities in access to reliable mental health care.

Emerging trends point toward multimodal integration, combining linguistic features with auxiliary data such as prosody, facial expressions, or physiological signals. Early research integrating LLM-driven text analysis with multimodal input (e.g., EEG data or speech patterns) shows substantial improvements in diagnostic precision [13], [17]. Future work should prioritize not only refining linguistic marker extraction via model-specific fine-tuning but also standardizing benchmarks for temporal and contextual evaluation [7].

In conclusion, the use of LLMs to analyze linguistic markers offers unprecedented dimensions of clinical insight, enabling personalized and scalable mental health diagnostics. While substantial gains have been made, including domain-specific pretraining and temporal tracking, concerns regarding cross-demographic biases and false positives must be addressed. By combining advances in language understanding with multimodal approaches, the next frontier in mental health diagnostics will involve moving beyond predictive capabilities to actionable, real-world solutions.

2.2 Predictive Screening and Early Risk Detection

The burgeoning role of Large Language Models (LLMs) in predictive screening and early risk detection for mental health interventions underscores their transformative potential in utilizing digital traces of human behavior. By analyzing user-generated content at scale, such as social media posts, blogs, and structured interview transcripts, these

models promise to identify early signs of mental health deterioration, enabling preemptive interventions before issues escalate to clinical levels. Through advanced natural language understanding, LLMs capture subtle linguistic, semantic, and contextual markers indicative of conditions like anxiety, depression, suicidal ideation, and other mental health challenges [4], [18].

One particularly impactful application of LLMs lies in analyzing unstructured content from social media platforms, where individuals openly express thoughts and emotions. Predictive models tailored to such data have demonstrated robust efficacy in identifying early risk factors for mental health conditions, often surpassing traditional diagnostic benchmarks. For instance, utilizing supervised learning combined with semantic feature extraction, researchers have achieved precise identification of Major Depressive Disorder (MDD) among Twitter users, with precision scores exceeding human diagnostic accuracy in certain studies [19], [20]. Moreover, hybrid approaches, such as ensemble classifiers or attention-based neural networks, have enhanced the accuracy of classifying mental health conditions like anxiety and PTSD by analyzing textual patterns and affective dynamics [21], [22].

Beyond detection, LLMs excel in risk stratification, enabling the categorization of individuals into tiers of urgency to prioritize care delivery. These systems often employ a combination of lexicon-based analysis and representational learning to rank individuals based on linguistic cues tightly correlated with distress markers, such as expressions of hopelessness, restlessness, or irrational thought processes [23], [24]. In particular, leveraging standardized metrics like the Patient Health Questionnaire-9 (PHQ-9), certain models align tweet-level linguistic indicators with depression severity, achieving strong concordance with clinician-annotated benchmarks and powering scalable pre-diagnostic assessment tools to inform healthcare priorities [25].

Expanding into structured conversational contexts, including psychometric assessments and diagnostic interviews conducted via conversational agents, LLMs have shown promise in automating symptom mapping to standardized diagnostic tools like PHQ and PCL. By leveraging adaptive prompting techniques, such systems infer clinical markers from user responses, enabling longitudinal tracking of symptoms and reducing clinician workloads [26]. Furthermore, fine-tuned models trained on domain-specific corpora provide personalized insights and generate interpretations that align with clinical standards, bridging the gap between automated prediction and practitioner guidance [5], [27].

Despite these advances, significant challenges remain. Predictive disparities arising from bias in training datasets undermine diagnostic fairness, often disproportionately impacting marginalized demographics and cultural groups [11], [28]. Additionally, the reliance on publicly available social media data raises privacy concerns and questions about ethical boundaries, particularly when risk detection occurs without explicit consent [11], [29]. False positives, which may result from exaggerated sensitivity in screening tools, risk straining healthcare systems and inducing unwarranted stress in low-risk individuals, while false negatives could delay critical interventions for those in need [30]. Addressing these obstacles will require fairness-conscious training protocols and the integration of multimodal signals—such as those derived from physiological indicators like EEG or vocal stress markers—to enhance diagnostic reliability [13], [31].

Looking ahead, advancing multimodal integration and fine-tuned instruction models represents a critical pathway toward more scalable, accurate, and context-aware early diagnostics. The intersection of LLMs with psychometric paradigms has the potential to expand preventive mental healthcare, but success will depend on addressing current limitations through interdisciplinary collaboration, ethical frameworks, and rigorous clinical validation protocols [32], [33]. Embedding predictive systems within a well-regulated ecosystem will enable LLMs to act as transformative tools in reducing the burden of mental health crises at a population level.

2.3 Ethical and Methodological Challenges in Diagnostics

The integration of large language models (LLMs) into mental health diagnostics offers unprecedented potential for early detection and assessment of mental health conditions through advanced natural language processing (NLP) capabilities. However, their deployment raises intricate ethical and methodological challenges that must be addressed to ensure fairness, reliability, and clinical utility. These challenges span biases in training data, limitations of generalizability, and risks of misdiagnoses, each carrying significant ramifications for both individual users and broader public health initiatives.

A core challenge lies in dataset bias, which can significantly affect the accuracy and fairness of diagnostic predictions. Mental health data used to train LLMs often exhibits demographic imbalances, with underrepresentation of minority groups, low-resource languages, or culturally distinct expressions of mental health symptoms. For example, linguistic markers of mental illness gleaned from Western-centric contexts may fail to generalize to non-Western populations, potentially leading to inequitable diagnostic outcomes [14]. Furthermore, demographic biases entrenched in historical data can exacerbate existing disparities in care. Studies have shown that biases related to race, age, and socioeconomic status introduce differential misclassification rates, particularly for marginalized groups [34]. These biases not only compromise fairness but may also erode trust among users from underrepresented groups, a critical issue when detecting sensitive conditions like depression or suicidal ideation.

Another methodological hurdle is the issue of generalizability, as diagnostic models trained on specific datasets often fail to perform consistently across diverse contexts. For instance, models that achieve high accuracy in detecting depression on a social media platform like Reddit may falter on platforms such as Twitter or Weibo, where linguistic conventions and user behaviors differ [18], [35]. This limitation is compounded by the dynamic nature of human language, where cultural shifts continuously reshape the manifestation of mental health-related language [36]. The necessity of robust cross-domain adaptations and continual

model fine-tuning thus emerges as a key methodological concern.

A further complication is the high risk of misdiagnoses, including both false positives and false negatives. False positives, where a healthy user is inaccurately flagged as at-risk or clinically significant, can lead to unwarranted emotional distress, stigma, or unnecessary engagement with mental health services. On the other hand, false negatives, failing to identify those at critical risk, may delay potentially life-saving interventions. Such errors are particularly concerning in applications detecting suicide risk, where prompt action is crucial. For example, models trained to detect suicidal ideation often struggle to reconcile semantic ambiguity in human expression, resulting in suboptimal performance [22], [37]. Temporal factors, such as shifts between acute episodes and periods of stability, further complicate the consistency of model predictions, underscoring the limitations of static text models in dynamic mental health contexts [30].

Ethical complexities also surround the detection of highly sensitive conditions like suicidality, where the consequences of detection—or nondetection—extend beyond misdiagnosis to privacy violations, agency erosion, and potential misuse of data. This is especially critical given the need to distinguish between passive signals of distress versus actionable crises. The ethical dilemma of intervening based on automated assessments must be tightly managed to balance proactive care against potential misuse or overreach, especially in non-clinical settings [29]. Moreover, models often produce opaque results, making it difficult to explain the reasoning behind their diagnoses, which can further inhibit user trust and clinician adoption [38].

Addressing these challenges necessitates focused methodological advancements and ethical foresight. Future work could prioritize diversification of training datasets through targeted inclusion of minority populations, culturally varied language patterns, and low-resource contexts [39]. Efforts to harmonize LLMs across multiple platforms or languages, leveraging techniques such as domain fine-tuning or multimodal data integration, may also enhance generalizability and robustness [40]. Finally, to safeguard against misdiagnoses, implementing human-in-the-loop systems and designing explainability-driven models could help align automated outputs with clinical best practices while maintaining accountability [41]. Rigorous evaluation frameworks integrating sensitivity-specificity trade-offs, subgroup performance metrics, and ethical safeguards will be essential for responsible deployment [42].

Ultimately, while LLMs hold immense promise for advancing mental health diagnostics, their ethical and methodological shortcomings must be rigorously addressed to ensure equitable, reliable, and transparent mental health care solutions.

2.4 Multimodal Approaches for Enhanced Detection

The integration of large language models (LLMs) with multimodal data sources represents a promising frontier in enhancing the early detection and diagnosis of mental health conditions. Mental health is inherently complex, involving intricate interactions between verbal expressions, nonverbal behaviors, and physiological states. While traditional text-based analysis proficiently captures linguistic markers, it often overlooks complementary dimensions such as speech characteristics, visual cues, and biometrics. Multimodal approaches, by combining textual data with audio, video, and physiological signals, provide a more comprehensive framework for assessing psychological states and diagnosing disorders. This subsection examines the advancements, capabilities, and limitations of LLM-driven multimodal diagnostic systems, emphasizing their transformative potential while aligning with preceding discussions on challenges in generalizability and validation.

A notable innovation within multimodal systems is the integration of textual data with audio features such as pitch, speech prosody, and fluency [17]. Cognitive and emotional disturbances often manifest through speech patterns, enabling the detection of mental conditions such as depression and post-traumatic stress disorder (PTSD). For instance, depressive states may correlate with slower speech and hesitations, while anxious individuals often exhibit heightened pitch or vocal tension. Studies combining natural language processing (NLP) with acoustic modeling have demonstrated significant gains in diagnostic sensitivity and specificity, particularly with LLMs fine-tuned for conversational settings [17], [43].

In addition to audio data, visual information significantly enriches diagnostic capabilities. Advances in computer vision techniques now enable the analysis of facial expressions, micro-expressions, and gaze patterns in conjunction with textual data from LLMs. Subtle shifts in facial muscle movements and expressions can act as reliable indicators of emotional dysregulation, complementing findings from verbal and acoustic features. For example, models leveraging 3D facial expression analysis have proven effective in assessing depressive symptom severity, with results aligning closely with clinical standards [17]. However, capturing high-quality, real-time visual data in naturalistic or uncontrolled settings remains a technical barrier, potentially limiting scalability and broader adoption.

Physiological signals such as heart rate variability (HRV), electroencephalograms (EEG), and galvanic skin responses add yet another layer to multimodal systems. These signals represent neural and autonomic responses directly associated with emotional and cognitive states, offering objective markers for mental health assessment. Integrating physiological data with LLM-driven textual analysis enables a shift from self-reports to measurable biomarkers, paving the way for precise, real-time diagnostics. For example, EEG-derived features combined with linguistic insights from LLMs have shown higher predictive accuracy for conditions like depression and anxiety disorders [44]. However, designing practical multimodal systems requires overcoming challenges related to data collection, user comfort with wearable devices, and standardization of physiological datasets across populations.

The convergence of multimodal data sources brings considerable advantages—enriched diagnostic inputs, reduced dependence on subjective self-reporting, and improved model robustness across diverse scenarios. However, this synthesis comes with its share of challenges. Training LLMs to process multimodal data demands large-scale, synchronized, and annotated datasets, which are still sparse for many mental health conditions. Furthermore, these systems

introduce computational complexity and latency concerns, necessitating innovations in resource-efficient architectures. Multimodal models are also susceptible to overfitting, particularly when datasets are small or imbalanced, underscoring a critical need for standardized benchmarks and rigorous validation protocols [45].

Emerging trends demonstrate the potential of hybrid architectures that combine pre-trained LLMs with task-specific models specialized in audio, visual, or physiological domains [17]. Enhanced multimodal datasets such as DAIC-WOZ and D^4 have contributed to iterative improvements in both model accuracy and explainability [43], [46]. Explainable AI methods further support these advances by elucidating which features or modalities influence a model's predictions, fostering stakeholder trust and aiding clinical integration [10].

Looking ahead, real-time multimodal systems integrating LLMs with multi-sensor data streams could revolutionize mental health diagnostics, especially in underserved or remote environments. Achieving this vision will require multidisciplinary collaboration to address gaps in dataset diversity, computational efficiency, and deployment scalability while maintaining ethical safeguards. Efforts to align these systems with the principles of equity and privacy are critical to ensure responsible adoption without perpetuating biases or exposing sensitive user information [47], [48].

By bridging the gap between fragmented diagnostic tools and holistic assessments, multimodal LLM-driven systems offer a transformative path forward, reinforcing the broader themes of innovation and validation discussed in adjacent subsections. As research continues to mature in this area, these technologies hold the potential to democratize access to comprehensive mental health evaluations, significantly advancing early diagnostics, interventions, and preventative care.

2.5 Real-World Validation and Benchmarking of LLM-based Diagnostics

Real-world validation and benchmarking of large language models (LLMs) in mental health diagnostics are critical to translating their demonstrated potential into practical, clinically reliable tools. This process involves rigorous evaluation across diverse dimensions such as diagnostic accuracy, sensitivity and specificity, longitudinal effectiveness, real-world variability, and alignment with existing clinical standards. This subsection delves into the methodologies and metrics employed for such validations, highlights current evidence and case studies, and identifies key challenges shaping the future of this domain.

Evaluation in real-world settings often begins with standardized benchmarks that align computational outputs with established mental health metrics. These benchmarks commonly leverage clinical instruments such as the Patient Health Questionnaire (PHQ-9), Columbia Suicide Severity Rating Scale (C-SSRS), or diagnostic frameworks derived from the Diagnostic and Statistical Manual of Mental Disorders (DSM-5). For example, systems augmenting PHQ-9 criteria with LLM capabilities have shown improved generalizability across datasets by grounding predictions in clinically relevant symptoms [24], [49]. Such benchmarks not only enhance validity but facilitate comparisons across different diagnostic tasks, such as depression severity estimation, suicide risk stratification, or anxiety classification.

Clinical trials and real-world applications are necessary to assess the robustness of LLMs when confronted with naturalistic data variability. Successful examples include the deployment of multimodal systems combining linguistic features with audio-video inputs to predict depression severity, outperforming isolated approaches reliant solely on textual data [17], [50]. Similarly, trials utilizing structured datasets like Self-reported Mental Health Diagnoses (SMHD) have demonstrated the efficacy of hybrid neural networks, such as SBERT-CNN combinations, for parsing semantic patterns in complex social media language [14], [51]. These studies underscore that fine-tuned LLMs often perform equivalently to or better than human clinicians in specific diagnostic tasks, achieving accuracy ranges of 80-85

One of the emerging trends in validating LLMs involves longitudinal monitoring and iterative model refinement. Systems integrated within clinical workflows are updated based on patient outcomes, enabling greater alignment with real-world context. For instance, temporal annotation of diagnosis statements has proven effective in assessing shifts in mental health conditions over time, allowing early warning markers to be identified and continuously fine-tuning models toward precision care [30]. Post-deployment monitoring frameworks analyzing user behavior have also highlighted risks of model hallucinations, over-reliance, and diagnostic inconsistencies, emphasizing the necessity of ongoing evaluation protocols [52].

The trade-offs in model performance represent another challenge in benchmarking. While transformer-based architectures such as RoBERTa and BERT have demonstrated state-of-the-art performance in tasks like stress and depression classification, their high resource demands limit usability in some clinical settings [16], [53]. Conversely, prompt-based systems leveraging computationally efficient techniques, such as bag-of-words classifiers, can approximate these results with significantly reduced resource requirements, albeit with trade-offs in granularity and complexity [54]. The tension between scalability and diagnostic precision remains pivotal in benchmarking discussions.

Future directions in this field must emphasize the integration of multimodal inputs—encompassing text, physiology, and non-verbal cues—to enhance diagnostic accuracy and holistic assessment [31], [55]. Efforts must also focus on culturally sensitive validation protocols to address biases and ensure equitable performance across diverse populations [13], [56]. These advances will require interdisciplinary collaboration among technologists, clinicians, and ethicists, supported by robust datasets and transparent evaluation standards.

In summary, real-world validation and benchmarking constitute the linchpin for adapting LLMs into effective diagnostic tools for mental health. Although current approaches have demonstrated considerable promise, sustained progress depends on a balance between methodological rigor, contextual adaptability, and ethical accountability.

3 Conversational Agents and Virtual Counseling

3.1 Core Capabilities of Conversational Agents in Mental Health

Conversational agents powered by large language models (LLMs) have emerged as a promising frontier in mental health interventions, demonstrating capabilities that extend beyond simple information retrieval to delivering nuanced, context-aware, and empathetic conversational support. Rooted in advancements in natural language processing (NLP) and deep learning, these agents leverage fine-tuned LLM architectures to support individuals facing psychological challenges. This subsection examines the technological advancements underpinning these capabilities, focusing on multi-turn dialogue management, empathy simulation, contextual awareness, and personalization, while highlighting the associated strengths, limitations, and emerging trends.

A defining feature of conversational agents for mental health is their ability to sustain coherent, multi-turn dialogues. Effective dialogue management ensures that these systems can maintain context over extended interactions, allowing for deeper, more meaningful exchanges. Modern LLMs, such as those adapted for healthcare contexts, like ClinicalGPT and Med42-v2 [57], [58], achieve this through memory mechanisms embedded in their architectures, such as attention-based transformers. These mechanisms preserve the user's conversational history, enabling agents to respond in ways that reflect prior input and convey continuity. This is particularly important for mental health scenarios, where user trust and engagement are fostered by the system's understanding of nuanced, long-form communication.

Empathy and emotional response simulation mark another critical capability of LLM-driven conversational agents. Effective empathy has been cited as a cornerstone of mental health support, helping users feel understood and validated. Fine-tuning LLMs with datasets annotated for empathic communication, such as the approaches highlighted in "A Computational Approach to Understanding Empathy Expressed in Text-Based Mental Health Support," equips these systems to generate responses with appropriate emotional tones [59]. Additionally, innovative frameworks like RESORT enable systems to assist users in emotional regulation by implementing evidence-based psychological strategies, such as cognitive reappraisal [60]. However, challenges remain in calibrating the authenticity of emotional responses. Users may detect inauthenticity, particularly in scenarios requiring deep empathy, which underscores the limitation of algorithmically generated emotions compared to human interaction.

Contextual awareness further enhances the effectiveness of conversational agents in mental health. By integrating syntactic and semantic analysis, models like Psy-LLM dynamically interpret user input to adapt responses that align with the user's emotional and thematic concerns [10]. These systems utilize reinforcement learning approaches, often guided by extensive domain-specific data, to refine their responses for complex contexts, such as trauma support or stress management. Advanced systems, like CBT-LLM, trained with cognitive behavioral therapy (CBT) principles, illustrate how structured intervention strategies can be embedded into conversational frameworks to assist users in reframing negative thought patterns [8]. Despite these advancements, addressing domain-specific nuances, such as cultural variations in psychological expression, remains a pressing challenge.

Dynamic content generation facilitates personalization, a hallmark of effective mental health interventions. Through fine-tuned models such as Mental-Alpaca and Mental-FLAN-T5, conversational agents can integrate user-provided historical data—ranging from previous dialogues to metadata signals—to generate customized advice or therapeutic prompts [5]. Such personalization enhances user satisfaction and therapeutic relevance. However, implementing personalization raises concerns about privacy and data security, particularly as sensitive mental health information is inherently vulnerable to misuse if not protected rigorously [11].

Despite their significant potential, trade-offs and limitations characterize these conversational agents. Issues such as providing overly generic or "hallucinated" responses, inaccuracies in mental health-related content, and biases against underrepresented demographics persist [61]. Moreover, while systems like DISC-MedLLM have demonstrated improved alignment with real-world consultation tasks, user over-reliance on automated systems at the expense of seeking human clinicians poses ethical risks [62].

Emerging trends suggest promising directions for future research. Multimodal integration, combining text, audio, and facial expressions, as seen in ongoing research incorporating EEG data [13], could enhance contextual depth and diagnostic accuracy. Additionally, there is increasing interest in frameworks that incorporate explainability and iterative improvement, allowing users and clinicians to evaluate the transparency of system recommendations [63].

In sum, LLM-driven conversational agents hold transformative potential for mental health support by blending dialogue coherence, empathy, contextual analysis, and personalization. However, achieving the delicate balance between technological sophistication and ethical responsibility remains a critical focus area, requiring further advancement in model alignment, evaluation, and safety protocols.

3.2 Applications in Evidence-Based Therapeutic Approaches

Large Language Models (LLMs)-powered conversational agents have emerged as promising tools for delivering structured mental health interventions by emulating principles of evidence-based therapeutic approaches. These systems, designed to integrate the core techniques of cognitive behavioral therapy (CBT), dialectical behavior therapy (DBT), and motivational interviewing (MI), offer scalable, personalized, and context-aware support for individuals seeking mental health assistance. This subsection explores their applications, highlights methodological innovations, and discusses trade-offs and challenges in effectively deploying these approaches within digital counseling frameworks.

Cognitive behavioral therapy (CBT) is one of the most widely implemented frameworks in LLM-powered interventions. These agents excel at facilitating CBT techniques

such as cognitive reframing and guided questioning, assisting users in identifying and reshaping maladaptive thought patterns into healthier cognitive schemas. Automated implementation of structured CBT-based interactions allows users to engage in exercises designed to reframe intrusive thoughts into constructive perspectives. For instance, studies presented in [64] and [65] illustrate the success of LLM-guided exercises in aligning user interaction with psychotherapy frameworks such as the Patient Health Questionnaire-9 (PHQ-9), fostering both emotional alignment and engagement. However, ensuring consistency and clinical rigor across diverse and dynamic user sessions remains a significant challenge, especially as these systems scale for widespread accessibility.

Whereas CBT concentrates on cognition, dialectical behavior therapy (DBT) broadens the therapeutic scope by emphasizing emotional regulation, mindfulness, and distress tolerance. LLMs have demonstrated potential in simulating DBT practices through scenario-based engagements that teach users practical coping mechanisms during crises. For example, DBT-based interventions deployed via conversational agents often guide users through mindfulness exercises or adaptive skill-building dialogues aimed at reducing emotional overwhelm. Reinforcement learning mechanisms further ensure these systems adapt to users' emotional tones, aligning with findings from [66]. Despite these advancements, DBT's interactive nature poses distinct challenges. Maintaining therapeutic safety and ensuring emotionally sensitive, dynamic conversations—especially in instances of acute distress—require capabilities not yet fully realized by current LLM systems.

Motivational interviewing (MI), a dialogic approach designed to foster behavioral change, also benefits from LLM-powered conversational systems. Through reflective listening, personalized encouragement, and guidance in navigating ambivalence, these systems prove effective in supporting users uncertain about committing to therapeutic changes. Platforms employing MI techniques have shown success in enhancing users' self-determination and motivation for improvement, as highlighted in [67]. However, MI's nuanced relational dynamics, which hinge on trust and the management of complex interpersonal cues, may be oversimplified in LLM-driven approaches, posing a risk of reduced authenticity in user interactions.

Across these therapeutic frameworks, an overarching strength of conversational agents lies in their ability to integrate psychoeducation alongside interventions, empowering users with a clearer understanding of their mental health. For example, frameworks discussed in [64] effectively combine psychoeducation modules with interactive exercises, bridging users' knowledge gaps and fostering agency. However, excessively generic or surface-level interventions can risk disengaging users, especially if they are not contextualized within individualized therapeutic goals and psychological frameworks.

Despite these promising developments, limitations and trade-offs remain critically important to address. The lack of genuine human empathy in LLM-powered agents limits their ability to build deep trust and connection in high-stakes mental health scenarios [28]. Additionally, concerns persist over the potential reduction of therapeutic interventions into formulaic scripts, detracting from the nuance and personalization required for effectively addressing diverse user needs. Ethical challenges—such as harmful outputs, user over-reliance on automation, and safety concerns in detecting or responding to escalatory scenarios (e.g., suicidal ideation)—necessitate robust monitoring tools and ethical safeguards, as emphasized in [68].

Emerging research underscores the potential of integrating multimodal capabilities, such as combining text-based dialogues with physiological or vocal data, to strengthen therapeutic models. For example, findings from [31] suggest that coupling linguistic insights with multimodal data enhances systems' abilities to holistically interpret emotional states and tailor interventions accordingly. In parallel, instruction-finetuned LLMs customized for evidence-based therapeutic domains [5] propose new directions for aligning LLM outputs more closely with established psychotherapeutic principles, ultimately delivering more precise and effective care.

In conclusion, the application of LLMs in evidence-based therapeutic approaches represents a significant step forward in democratizing access to scalable and affordable mental health resources. By operationalizing frameworks such as CBT, DBT, and MI, these systems hold transformative potential to complement existing clinical care structures. Nonetheless, achieving fidelity to therapeutic frameworks, meeting ethical obligations, and addressing gaps in empathy and contextual adaptation will require sustained interdisciplinary research and innovation. Advancing personalization, safety, and cultural sensitivity will be essential for enabling the widespread adoption of LLM-based interventions in real-world mental health scenarios.

3.3 Improving Accessibility and Anonymity in Mental Health Support

The integration of large language model (LLM)-based conversational agents into the domain of mental health has demonstrated transformative potential in addressing some of the most persistent challenges in providing equitable access to care, particularly for underserved and stigmatized populations. By offering scalable, context-sensitive, and anonymous support, these agents serve as innovative tools to bridge gaps in mental health accessibility and mitigate barriers such as geographical restrictions, social judgment, and resource limitations.

Scalability is one of the most compelling attributes of LLM-driven mental health support systems. These systems harness LLMs' ability to engage in high-quality multi-turn conversations, delivering real-time support to individuals across vast geographies, including low-income and rural areas where trained professionals are in short supply [29], [69]. Unlike conventional mental health services that are limited by human resources, conversational agents offer a scalable alternative, operating continuously to meet the demand for support. Notably, the lightweight architectures of some LLM implementations have enabled their efficient operation on low-resource devices, making them suitable for deployment in regions with minimal computational infrastructure [70], [71], [72]. This capacity to scale mental health services emerges as an unprecedented opportunity

to improve access for individuals globally, particularly in regions with limited mental health infrastructure.

Anonymity is another pivotal feature offered by LLM-based conversational agents, and one that directly addresses the stigma associated with mental health care. Traditional services often deter individuals from seeking help due to the fear of judgment or exposure of vulnerabilities. Virtual counseling via conversational agents, however, provides a confidential platform for individuals to share their experiences freely and without the anxiety of social repercussion [29]. Furthermore, platforms powered by advanced LLMs like ChatGPT and GPT-4 have repeatedly demonstrated the ability to engage in supportive and empathetic interactions, fostering trust with users [5], [15]. Specifically, such agents have shown promise in sensitive contexts, such as providing tailored emotional support to marginalized groups, including queer youth, while respecting their unique cultural and social identities [73].

Multilingual and culturally sensitive designs of LLM-based systems further enhance their accessibility and inclusiveness. The ability of multilingual models to engage users in their native languages ensures that these systems are not constrained by linguistic barriers, which often exclude non-native speakers from accessing traditional mental health care [14], [72]. Furthermore, fine-tuning LLMs with culturally contextualized datasets allows these agents to respond sensitively to diverse cultural norms and expressions of mental health distress. However, challenges remain in addressing underrepresented languages and dialects sufficiently, as many LLMs are still biased toward languages with greater digital resources [14], [34].

Despite these strengths, scaling LLM-based accessibility solutions is accompanied by trade-offs. For one, the quality of interaction is still limited by the inherent inability of LLMs to replicate authentic human empathy during complex therapeutic exchanges [29]. Additionally, ethical concerns arise regarding the privacy and security of user data, especially in cases where personal narratives or sensitive health information is shared during interactions. These concerns underscore the necessity of robust encryption techniques and differential privacy methodologies to ensure the anonymity and security of user data [5]. It is also vital to address biases in model training data that may inadvertently perpetuate inequities in mental health support among different demographic groups [36].

Looking forward, incorporating multimodal capabilities into LLM-powered conversational agents may further revolutionize mental health accessibility. These advancements could enable models to integrate text with other modalities, such as speech and biometric data, to provide more personalized and holistic support [40]. Research must also continue toward improving the interpretability and transparency of LLM outputs, which will both increase user trust and align generated content more closely with evidence-based therapeutic practices [5]. By addressing these challenges proactively, LLM-based systems can fulfill their promise in transforming mental health care into a more inclusive, equitable, and anonymous service for populations across the globe.

3.4 Limitations and Ethical Considerations in Virtual Counseling

The integration of large language model (LLM)-based conversational agents into virtual counseling presents both immense opportunities and significant challenges, particularly within the high-stakes and sensitive context of mental health care. While these systems demonstrate promise in providing scalable, flexible, and context-aware interactions, their limitations highlight critical ethical and practical considerations. This subsection explores key challenges, including emotional authenticity, the risk of harmful outputs, dependency on automated systems, and the delicate balance between automation and human oversight, as informed by state-of-the-art research.

One of the central challenges of LLM-powered conversational agents is their inability to replicate genuine human empathy, a critical element for building trust and rapport in therapeutic relationships. LLMs, fine-tuned to recognize emotional cues and deliver empathetic responses, can simulate understanding to some degree [67]. However, this simulation often lacks the depth and nuance required to address the complex and dynamic nature of human emotions. Achieving this kind of emotional authenticity would require not just linguistic capabilities but also multimodal integration, such as interpreting tone, speech prosody, and potentially even body language, which these systems currently lack [17]. Many users report that while these agents provide non-judgmental and supportive interactions, the absence of true comprehension and genuine recognition of individual complexity can foster a sense of artificiality, undermining the therapeutic experience [74]. This challenge is compounded by the inherent probabilistic nature of LLMs, which generate outputs based on prediction rather than understanding, further limiting their capacity for deep emotional engagement.

An equally pressing concern lies in the risk of hallucinations, where conversational agents may produce incorrect or even harmful outputs—an issue with heightened consequences in mental health scenarios. LLMs, including advanced models like GPT-4, are known to sometimes generate misinformation with unwarranted confidence, which can be misinterpreted by users as credible advice [75]. In high-stakes situations such as crisis interventions or when addressing suicidal ideation, these errors can carry severe repercussions [76]. Efforts to address this issue include the development of structured evaluation frameworks and domain-specific safety measures designed to minimize the risk of harmful outputs [68]. However, the unpredictable nature of hallucinations underscores the critical need for more robust safeguards and transparent operational protocols.

The potential for dependency on these automated systems poses another concern. While LLM-driven virtual counseling tools hold the potential to break down barriers such as stigma and resource constraints, there is evidence that vulnerable users might over-rely on these systems and delay seeking timely professional intervention [10]. This risk of dependency becomes especially problematic when these tools, while adept at handling lower-complexity tasks, struggle with nuanced therapeutic interventions, leaving

critical cases unaddressed [29]. Ensuring that these systems appropriately encourage users to pursue additional care when necessary remains an essential consideration in their design and deployment.

Finally, the ethical challenges inherent in balancing automation and human oversight are significant. Fully automated systems may fall short in reliably handling sensitive scenarios, such as crisis interventions, while hybrid approaches involving human review introduce scalability constraints. Potentially effective solutions include hybrid frameworks where LLM responses are monitored and supplemented by professionals in real-time, particularly in high-risk contexts [68]. Additionally, these systems can unintentionally amplify existing social biases or stigmas embedded in training data, including those related to gender and mental health diagnoses, necessitating fairness-aware frameworks and ongoing efforts to improve dataset diversity [77], [78]. Addressing these biases is critical to ensure that these solutions contribute to equitable and unbiased mental health support.

In summary, while LLM-based conversational agents hold transformative potential to enhance access to mental health care through scalability and inclusivity, their current limitations demand thoughtful and deliberate approaches to their deployment. Challenges such as achieving emotional authenticity, safeguarding against harmful outputs, mitigating user dependency, and balancing automation with human oversight must be carefully addressed to ensure safe and effective use. Interdisciplinary collaboration among technologists, clinicians, and ethicists is paramount to ensure that these systems align with the ethical principles of mental health care. Future refinements should emphasize multimodal capabilities, robust ethical frameworks, and rigorous evaluation standards to create responsible virtual counseling tools that meaningfully complement human-led therapeutic interventions.

3.5 Evaluation and Metrics for Conversational Agents in Mental Health

The evaluation and measurement of large language model (LLM)-powered conversational agents in mental health applications remain critical for ensuring their efficacy, safety, and usability. Rigorous and continuous evaluation frameworks are paramount, given the sensitive nature of clinical interventions and the ethical implications of deploying such tools at scale. This subsection explores key metrics, methodologies, and challenges in assessing these systems, with an emphasis on aligning evaluations with clinical and therapeutic benchmarks.

A comprehensive evaluation framework for conversational agents in mental health typically combines technical performance metrics, therapeutic efficacy indicators, and user-centered assessments. The therapeutic alliance—the degree to which users feel understood and supported during interactions—is one of the most pivotal measures, often assessed through user satisfaction surveys and post-session feedback loops [15], [79]. While traditional metrics like coherence, fluency, and response accuracy address the linguistic quality of generated outputs, mental health-specific evaluations also prioritize empathy and emotional alignment, often formalized as "empathy and comfort scores." For example, studies have proposed using psycholinguistic analyses and human annotations to score model responses on their ability to convey empathy effectively and foster user comfort [29], [79].

Outcome-based efficacy metrics focus on the tangible benefits users derive from engaging with these agents. These can include reductions in self-reported symptom severity, increased therapeutic commitment, or measurable improvements in mood or well-being. For instance, incorporating clinically validated scales such as the Patient Health Questionnaire (PHQ-9) into model interactions enables quantitative tracking of user progress over time [27], [49]. Outcome measurement frameworks may also employ pre- and post-session comparisons to understand the interventions' direct psychological impact or rely on longitudinal methods to assess the model's role in sustained mental health improvement [17].

Technological advancements have introduced multimodal approaches to evaluation, blending textual, auditory, and even visual data to provide a holistic assessment of conversational agents. For example, the integration of acoustic and prosodic speech features with text-based analysis can significantly enhance the sensitivity of depression detection and emotional state tracking, as seen in multimodal attention networks [31], [80]. Furthermore, methods like temporal annotation allow for dynamic performance evaluations across the course of multiple user interactions, tracking shifts in therapeutic engagement or emotional patterns over time [30], [81].

Despite advances, existing methodologies face several challenges. A significant limitation involves the subjective variability in evaluating empathy and therapeutic alliance. Human annotations, while essential, are resource-intensive and prone to bias. To address this, emerging approaches involve leveraging automated psychometric tools, such as those enabled by large-scale transformer models, to score interactions based on predefined therapeutic constructs [9], [82]. Another challenge lies in the trade-offs between generalizability and specificity. Models trained on broad, diverse datasets often struggle to replicate the specialized skillsets required for sensitive mental health contexts. Fine-tuning these models with domain-specific data holds promise but remains computationally demanding and requires robust annotated corpora [10], [51].

An increasingly critical area of focus lies in bridging the gap between formal psychotherapy standards and automated evaluation methods. Current methods tend to overlook the nuanced ethical and safety considerations, such as preventing harmful outputs or maintaining professional boundaries in virtual counseling settings [29], [83]. Continuous evaluation frameworks, driven by real-world feedback and clinical validation studies, are necessary to mitigate these risks effectively. Iterative Post-Deployment Audit Frameworks (PDAFs), for instance, have been proposed to align conversational agents' predictive and therapeutic capabilities with clinical ground truths, ensuring their applicability in real-world scenarios.

Looking forward, ensuring multimodal integration, real-time monitoring, and adaptive feedback mechanisms will be key to advancing the evaluation landscape of conversational agents in mental health. Research efforts must

prioritize achieving interoperability between these agents and existing healthcare systems, while also fostering greater transparency in evaluation processes. As LLMs continue to evolve, the field will benefit immensely from interdisciplinary collaborations among technologists, clinicians, and ethicists to create safer, more effective tools for digital mental health care.

4 Enhancing Professional Practice in Mental Health

4.1 Automation of Administrative and Clinical Documentation

The automation of administrative and clinical documentation represents a pivotal application of large language models (LLMs) in mental health, addressing the pervasive challenge of clinician workload. By leveraging the advanced natural language processing (NLP) capabilities of LLMs, mental health practitioners can significantly reduce the time spent on routine documentation tasks while preserving the granularity and accuracy required for high-quality patient care.

LLMs have demonstrated the ability to convert unstructured conversational data from therapy sessions into structured, actionable documentation. For instance, tools powered by models such as GPT-4 and similar architectures are capable of summarizing therapy dialogues, extracting key clinical markers, and organizing content into standardized templates. These capabilities echo findings from studies where LLM-driven systems matched or even exceeded humans in clinical text summarization tasks, emphasizing their efficacy in generating patient reports and progress summaries [84]. Such automated processes allow clinicians to focus more on direct patient interaction, reducing administrative fatigue without sacrificing documentation quality.

A distinguishing feature of LLM-driven automation is the seamless integration of transcription and annotation capabilities. Advanced LLM systems can process audio recordings of therapy sessions and provide both verbatim transcriptions and selectively annotated highlights relevant to clinical decision-making. This dual functionality has been particularly impactful in contexts requiring the preservation of contextual nuance, such as documenting shifting patient affect or indications of distress. In comparative evaluations, LLMs like ClinicalGPT and GatorTron have consistently demonstrated superior performance in extracting clinical concepts and contextual summaries, showcasing their alignment with the complex demands of mental health documentation [57], [85]. By automating these workflows, LLMs alleviate the cognitive burden on providers while maintaining compliance with professional standards.

Key advantages of using LLMs for documentation extend beyond time savings. The scalability of this technology ensures broad applicability across diverse clinical scenarios, from individual psychotherapy to more complex multidisciplinary care settings. Models fine-tuned for mental health contexts, such as MentalBERT and CBT-LLM, have demonstrated a strong capacity to adapt to specialized terminologies and evidence-based intervention frameworks, enabling precise documentation tailored to various therapeutic modalities [2], [8]. Additionally, their multilingual capabilities empower clinicians to serve linguistically diverse populations, an imperative in globalized mental health care systems [86].

However, challenges persist in deploying LLMs for clinical documentation. One notable concern is the risk of inaccuracies stemming from hallucinated model outputs, particularly in high-stakes contexts where erroneous documentation may influence critical care decisions. Research evaluating GPT-3.5 and GPT-4 in healthcare settings highlights that while these models often generate linguistically accurate outputs, they struggle with domain-specific alignment and may propagate errors under certain conditions [61]. Moreover, ethical considerations such as data privacy and confidentiality must be rigorously addressed, as mental health records are among the most sensitive forms of clinical documentation. Techniques like differential privacy and on-device processing offer promising pathways to safeguard data, yet their integration remains underexplored [85].

Emerging research has also begun to probe the potential of LLM-assisted documentation as a collaborative tool rather than a fully autonomous process. For example, hybrid systems where clinicians interact with LLM-generated drafts, refining and contextualizing outputs, could bridge gaps between automation and human expertise. Studies suggest such frameworks can optimize the balance between efficiency gains and maintaining clinician oversight, ensuring reliable and ethically sound documentation practices [87].

Looking forward, future advancements in LLM-based clinical documentation could benefit from multimodal integration, combining text with speech, video, and even physiological data for comprehensive patient records. Such approaches would not only enhance the depth of documentation but also allow for richer longitudinal tracking of patient progress, a vital aspect of mental health services [13]. Furthermore, ongoing improvements in instruction tuning and domain-specific pretraining, such as efforts showcased in models like Psy-LLM and DISC-MedLLM, are likely to yield systems better aligned with professional standards and nuanced care delivery requirements [10], [62].

In conclusion, the application of LLMs to automate clinical documentation in mental health holds transformative potential. By reducing administrative burdens while enhancing documentation accuracy and scalability, these models can contribute to improved provider well-being and patient outcomes. However, careful attention to limitations and ethical safeguards is paramount to ensuring that this leap forward genuinely benefits mental health systems. Future developments should prioritize interdisciplinary collaboration and iterative refinement to address the nuanced demands of mental health care documentation comprehensively.

4.2 Decision Support for Evidence-Based Practices

Large language models (LLMs) are demonstrating transformative potential as decision support systems in mental health, serving as invaluable tools to facilitate evidence-based practices while complementing advancements in automation and professional training. Their ability to retrieve,

synthesize, and recommend evidence-based research in real time is reshaping the landscape of clinical decision-making. This subsection examines the mechanisms through which LLMs bolster evidence-based practices, highlights their strengths and limitations, and discusses ongoing challenges and future opportunities for advancement.

A key strength of LLMs in decision support lies in their unparalleled capacity to process and analyze vast volumes of clinical and research data, converting it into actionable insights. Leveraging architectures such as GPT and specialized biomedical models like Med-PaLM, these systems excel at distilling unstructured data—ranging from electronic health records (EHRs) to clinical guidelines and scholarly articles—into accessible, clinician-friendly formats. For example, LLMs can operationalize Patient Health Questionnaire (PHQ) scores, aggregate research on therapeutic modalities, and reconcile variances in clinical recommendations [69], [88]. Their contextualization capabilities further allow them to tailor outputs to individual patients' symptoms, demographic factors, and cultural considerations, fostering informed and personalized decision-making.

A notable application of LLM-based decision support is in providing personalized intervention recommendations. Advanced LLMs trained on mental health-specific datasets, such as SMHD and PRIMATE, demonstrate the ability to identify patterns in linguistic and behavioral data, enabling dynamic suggestions for diagnostic evaluations and therapeutic interventions. By integrating structured assessments like PHQ-9 or PCL-C within interactive platforms, these models not only assist clinicians in real-time evaluation but also adapt recommendations in light of a patient's mental health trajectory over time [19], [26]. However, challenges remain in ensuring these systems are inclusive and equitable, as their effectiveness can diminish when dealing with underrepresented conditions or populations with scarce data representation [35], [69].

LLMs further strengthen evidence synthesis by bridging the gap between mental health research findings and practical clinical application. Through fine-tuning on domain-specific data, LLMs are capable of summarizing complex therapeutic research, distilling relevant patterns from unstructured clinical narratives, and offering synthesis-driven guidance [35], [89]. Such capabilities are particularly valuable in multidisciplinary care environments, where clinicians benefit from rapid translations of emerging evidence into actionable insights. For instance, automated systems have shown efficacy in transforming raw therapy transcripts into thematically organized outputs, aiding both qualitative and quantitative decision-making [38], [88]. Nevertheless, the inherent opacity of LLM-generated outputs introduces challenges regarding explainability and the risks of generating unverified content, particularly when the stakes involve therapy recommendations or diagnostic judgements [28], [83].

Emerging trends in LLM applications show promise for enhancing these systems through multimodal integration, combining text with data streams, such as speech patterns, facial expressions, or physiological indicators like EEG. By incorporating acoustic and visual modalities, LLM-enhanced decision support systems could provide a richer, more holistic understanding of mental health presentations. For instance, advancements in analyzing spoken language and 3D facial expressions have shown potential in improving diagnostic accuracy for complex disorders such as PTSD and depression, which often encompass both verbal and non-verbal markers [13], [17]. These innovations signify a shift towards context-aware, multimodal platforms that can help clinicians capture the multi-faceted nature of mental health conditions.

Despite their transformative potential, several challenges impede the broader implementation of LLM-driven decision support. Risks related to biases in training data, inconsistencies in outputs, and concerns about cultural and demographic inclusivity require urgent attention. Research evaluating GPT-4, for example, has noted discrepancies in empathy levels and fairness when responding to diverse demographic subgroups, underscoring the need for fairness-aware fine-tuning and expanded representation in datasets [12], [28]. Furthermore, building trust with clinicians will require advancements in the explainability of model-generated recommendations, ensuring accountability in high-stakes mental health care decisions [36], [88].

As mental health services increasingly recognize the necessity of evidence-based decision support, LLMs have the potential to become indispensable tools in improving intervention quality and accessibility. Future directions should emphasize the integration of real-time feedback loops, multimodal data, and fairness-aware development methodologies to ensure these systems meet the nuanced demands of modern mental health care. By aligning cutting-edge technological innovations with the principles of ethical care, LLMs could significantly enhance the precision, adaptability, and inclusivity of mental health interventions, further solidifying their role as transformative allies in evidence-based practices.

4.3 Training Tools for Clinicians and Mental Health Students

The integration of large language models (LLMs) into the education and training of mental health professionals has introduced a paradigm shift toward immersive, scalable, and contextually adaptive learning environments. This subsection explores how LLMs are reshaping clinician education and the training of mental health students through the simulation of realistic patient scenarios, interactive therapeutic strategy training, and personalized learning pathways, offering a blend of pedagogical innovation and practical applicability.

One of the most impactful contributions of LLMs is their ability to simulate diverse and realistic patient interactions, supporting the development of diagnostic acumen and counseling skills in mental health professionals. By leveraging LLMs' capacity for multi-turn dialogue and context retention, educators and trainees can engage with virtual patients who mimic the speech patterns, emotional states, and cognitive distortions characteristic of individuals with psychological disorders. For example, models fine-tuned on mental health-specific datasets enable the creation of patient simulations exhibiting conditions such as depression, anxiety, or PTSD. These simulations can include nuanced linguistic markers like negative self-talk, abrupt sentiment

shifts, or fragmented narrative structures, which are commonly indicative of underlying mental illness [14], [30]. Such interactions provide a safe space for trainees to practice interviewing techniques, explore differential diagnoses, and apply evidence-based therapeutic frameworks without the risk of causing harm to real patients.

Interactive therapeutic strategy training constitutes another key contribution of LLMs to professional development in mental health. By generating dynamic dialogues that adapt to clinicians' therapeutic approaches, LLMs enable trainees to explore the efficacy of specific techniques, such as cognitive restructuring in Cognitive Behavioral Therapy (CBT) or reflective listening in Motivational Interviewing (MI). Advanced LLMs, particularly those augmented through fine-tuning or instruction-based learning paradigms, allow therapists to test and refine interventions in real-time, receiving immediate feedback on their chosen approaches [5]. Such a capability bridges a critical experiential gap in education, enabling richer hands-on practice outside of supervised clinical internships.

A particularly advantageous feature of LLM-powered training systems is personalization. By leveraging few-shot prompting and historical session data, these models can continuously adapt educational content to the trainee's proficiency and learning trajectory. For instance, an LLM could scaffold complex topics over time, starting with foundational principles such as identifying linguistic tone and sentiment before progressing to multifactorial diagnostic approaches involving socio-demographic and cultural considerations [32], [38]. Additionally, personalized question-generation frameworks grounded in diagnostic protocols, like ProKnow, offer structured teaching experiences specifically aligned to clinical guidelines, ensuring that trainees' learning objectives are met while aligning with established safety and ethical standards [41].

Notwithstanding these advancements, limitations persist within current LLM training frameworks. Emotional authenticity—essential to counseling and therapeutic rapport—remains a nuanced challenge. While LLMs can simulate empathetic dialogues, their responses often lack the genuine warmth and spontaneity that characterize human interaction. This limitation raises important questions about the adequacy of simulated training scenarios for preparing clinicians to navigate the emotive complexities of mental health practice [9]. Additionally, risks of systemic biases in LLM training data pose challenges to fairness and inclusivity when simulating patients from underrepresented demographics [29]. Mitigating these challenges will require ongoing efforts to align LLMs more closely with specialized clinical inputs, ensure bias-aware training pipelines, and integrate multimodal data sources such as audio and video signals for more robust context modeling [40].

Looking ahead, the role of LLMs in training mental health professionals is projected to extend beyond technical simulation toward deeper integration within the broader educational ecosystem. Future developments could include augmented reality tools that combine LLM-driven dialogues with virtual patient avatars, thereby offering trainees a multisensory experience of client interactions. Moreover, adaptive curriculum design informed by real-time performance analytics could enable LLM-based systems to serve as long-term mentors, guiding clinicians through advanced stages of their careers [5].

In conclusion, large language models represent a transformative tool for professional development in mental health, offering scalable, adaptive, and theoretically grounded training solutions. By addressing existing limitations and continuing to refine their clinical and pedagogical alignment, these technologies have the potential to revolutionize how mental health professionals are educated, ensuring that they enter the workforce with greater proficiency, confidence, and readiness to tackle the complexities of modern mental health care.

4.4 Integrated Workflow Enhancement

The integration of large language models (LLMs) into existing mental health workflows offers a transformational avenue to enhance clinical efficiency, reduce administrative burdens, and support evidence-based decision-making. To realize their potential, it is imperative to focus on ensuring compatibility with healthcare systems, fostering interdisciplinary collaboration, and addressing resistance to automation among mental health professionals. This subsection examines these aspects, presenting challenges and opportunities for seamless adoption.

A cornerstone of LLM adoption is their interoperability with electronic health record (EHR) systems, which serve as the backbone of modern mental healthcare. LLMs can complement EHR functionality by automating numerous complex processes, such as clinical documentation and predictive analytics. For instance, generative models like GPT-4 have shown proficiency in summarizing therapy sessions and producing detailed case notes that adhere to established clinical standards [5], [90]. Furthermore, domain-specific fine-tuning, as seen in frameworks like PsycoLLM, enables these systems to tailor outputs based on mental health-specific criteria, enhancing reliability and customization [91]. However, challenges remain, notably errors stemming from incomplete contextual understanding or hallucinated outputs. As such, clinician oversight is indispensable to ensure accuracy in these high-stakes environments [75].

Beyond individual workflows, LLMs offer considerable promise in streamlining collaboration within multidisciplinary mental health teams. For example, models like PsyLLM are designed to facilitate centralized communication by providing real-time insights gleaned from patient data and literature summaries, aiding in collaborative decision-making [10]. This functionality has the potential to close gaps between clinicians, social workers, and support staff by generating shared action plans or flagging urgent cases requiring immediate intervention. However, challenges such as redundant information outputs and misalignment in clinical goals may hinder adoption [9]. Tailoring these systems to specific roles and responsibilities within care teams by employing adaptive personalization, as demonstrated in Mental-FLAN-T5, could help navigate such issues and foster broader acceptance [5].

Nevertheless, resistance to LLM-driven automation persists among clinicians, largely driven by concerns about reliability, professional agency, and ethical accountability. Mental health professionals often express skepticism about

relying on automated systems in contexts requiring nuanced human judgment, particularly for empathetic interactions [92]. Addressing this hesitancy necessitates embedding explainability mechanisms within LLM-driven workflows. For instance, visualization tools like saliency maps or attribution techniques can improve transparency by illustrating the reasoning behind specific recommendations or summaries [93]. Structured feedback loops that iteratively refine outputs based on clinician input represent another avenue for enhancing trust and fostering co-adaptiveness between users and systems [68].

Emerging trends also suggest exciting opportunities to embed LLMs more deeply into broader healthcare ecosystems. For instance, integrating real-time multimodal data—such as text, speech, and physiological signals—has the potential to significantly enhance diagnostic precision and therapeutic interventions [17]. Combining these capabilities with wearable technology for continuous mood and behavior monitoring could provide clinicians with richer diagnostic contexts. However, these advancements also demand the implementation of robust data governance frameworks to safeguard patient privacy [94].

In summary, integrating LLMs into mental healthcare workflows offers transformative potential but hinges on delivering reliable, contextually relevant, and clinically aligned outputs. Strategies such as domain-specific fine-tuning, transparent system architectures, and clinician-centric design can mitigate barriers to adoption while amplifying the efficiency and quality of mental health services. Moving forward, research should prioritize developing standardized implementation guidelines and robust evaluation metrics to ensure consistent performance across diverse care settings. By addressing these challenges, LLMs can position themselves as indispensable tools in advancing global mental health practices while maintaining ethical and professional accountability [11], [95].

4.5 Enhancing Client-Clinician Interactions through Augmentation

The integration of large language models (LLMs) into mental health practice offers profound potential to augment client-clinician interactions by providing real-time insights, enhancing therapeutic communication, and personalizing interventions. These advancements not only optimize session quality but also enable clinicians to respond more effectively to the nuanced needs of their clients, fostering deeper therapeutic alliances.

One of the primary applications of LLMs in client-clinician interactions is real-time sentiment and tone analysis. By leveraging natural language processing capabilities, LLMs can analyze clients' linguistic patterns during sessions, identifying shifts in emotional tone or distress levels in real-time. For instance, tools that measure sentiment gradients or detect markers of heightened anxiety or sadness during live conversations allow clinicians to adjust their responses dynamically and ensure empathetic engagement. Such advancements align with findings that emphasize the efficacy of linguistic analysis for identifying mental state indicators [4], [96]. An extension of this approach includes incorporating multi-modal integration where vocal intonation and micro-expressions are linked to text-based sentiment detection, providing a richer analysis of client emotions [17].

Cultural and linguistic adaptability represents another critical innovation facilitated by LLMs. These models can be fine-tuned to understand and interpret the linguistic nuances of clients from diverse backgrounds, enabling clinicians to deliver culturally responsive care. For instance, multilingual models trained on region-specific idioms and culturally sensitive speech styles can better capture underlying mental health concerns that traditional approaches might overlook. This flexibility is particularly vital in treating clients from underrepresented linguistic groups, where conventional datasets are often inadequate [14].

LLMs are also instrumental in augmenting therapeutic communication by generating tailored interventions during sessions. For example, an LLM could suggest clinically validated follow-up questions or interventional statements based on the flow of the session, effectively enriching the clinician's therapeutic repertoire. Cognitive behavioral therapy (CBT)-based interventions generated in this manner mirror structured therapeutic principles, offering suggestions grounded in robust evidence bases such as the PHQ-9 or DSM-5 frameworks [24], [49]. Moreover, systems that prioritize explainability, such as those designed to highlight the reasoning behind suggested client-linguistic response pairings, enhance the clinician's trust in such augmentative technologies [27].

While the promise of LLM-enhanced interaction is significant, several challenges remain. For instance, reliance on real-time outputs raises concerns of over-dependence and risks associated with inaccuracies, such as generating misleading interpretations or culturally inappropriate responses—problems exacerbated by biases in training data [16], [83]. Furthermore, the models' ability to navigate complex emotional landscapes, such as distinguishing between figurative and literal language in client speech, underscores the need for systems equipped with robust figurative language understanding frameworks [24]. These limitations highlight the importance of iterative evaluation and clinician oversight during deployment.

Future advancements could focus on integrating LLMs with multi-modal data to create holistic interaction platforms. By combining textual analysis with biometric indicators such as EEG signals or prosodic features in speech, such systems could offer unprecedented precision in detecting and responding to mental health indicators [13], [31]. Additionally, as LLMs continue to evolve, it will be crucial to prioritize methods for explainability and accountability to ensure that augmentative outputs align with clinical best practices and ethical guidelines [66], [69].

Overall, LLMs present a radical opportunity to deepen the client-clinician dynamic through their capacity to provide tailored, adaptive, and evidence-based augmentation. However, their effective implementation requires addressing data and deployment challenges while maintaining human oversight to preserve the integrity and empathy central to mental health care. By fostering interdisciplinary collaborations and advancing technological refinement, the field can continue to unlock novel pathways to improve session quality and client outcomes.

4.6 Addressing Limitations and Ethical Implementation

The integration of large language models (LLMs) into professional mental health practices holds immense promise, but it also necessitates an in-depth examination of limitations and ethical considerations to facilitate responsible deployment. This subsection systematically explores these challenges, identifying key barriers to adoption, evaluating current mitigation strategies, and proposing a forward-looking framework for ethical and effective implementation—complementing the preceding discussions on client-clinician interactions and bridging toward future applications.

A foundational limitation of LLMs in mental health is the potential violation of data privacy and confidentiality due to their reliance on large datasets for training. Mental health data is particularly sensitive, and the leakage of private information, whether through direct exposure or reconstructed narratives, carries significant ethical and legal risks. Anonymized data often remains vulnerable, as linguistic or contextual cues can lead to re-identification, raising important questions about the robustness of current anonymization techniques [5], [10]. Addressing these concerns necessitates integrating cutting-edge data minimization and differential privacy methodologies throughout both training and deployment phases [5]. Federated learning approaches, which enable data processing on local user devices rather than centralized servers, also offer a promising pathway to reduce privacy risks in this sensitive domain [10].

Another persistent challenge relates to inaccuracies, biases, and the risk of perpetuating stereotypes in LLM outputs. Since LLMs rely on patterns within their training data, they can inadvertently encode and reproduce cultural, gender, and socioeconomic biases, potentially leading to inequities in mental health support [11]. For example, individuals belonging to underrepresented demographic groups may encounter less accurate or less empathic responses from LLMs, as observed in recent studies [12]. Addressing these disparities involves augmenting LLMs with fairness-aware training datasets and counterfactual data augmentation techniques, though these methods remain imperfect [1]. Complementary mechanisms such as transparent bias auditing, where controlled demographic prompts analyze inequities, can further preemptively identify and rectify potential bias during deployment [11].

The opacity inherent in LLM architectures compounds these limitations, as black-box models do not readily offer interpretable outputs, making it difficult for clinicians to validate recommendations or diagnoses against clinical standards [97]. Incorporating explainable AI techniques, such as saliency maps and chain-of-thought reasoning outputs, can address this gap by offering enhanced transparency for clinicians [9]. Moreover, aligning these techniques with established clinical diagnostic benchmarks, such as the PHQ-9 guidelines, increases trustworthiness and generalizability in high-stakes mental health contexts [49].

Accountability represents yet another pivotal ethical challenge, particularly in determining responsibility for adverse outcomes in LLM-augmented clinical settings. The ambiguity surrounding the division of responsibility between human practitioners and LLMs may lead to trust gaps and legal liability concerns [98]. Establishing robust clinician oversight for LLM-generated outputs ensures that critical decision-making remains under professional jurisdiction. Simultaneously, transparent communication about the limitations of these tools—via ethical informed consent processes—can help manage expectations for both clients and clinicians [69]. Involving diverse stakeholders, including clinicians, ethicists, and patients, in the development pipeline fosters participatory design and strengthens shared accountability [99].

Iterative evaluation and ongoing refinement of LLMs remain crucial in navigating the dynamic, high-stakes landscape of mental health care. Post-deployment, real-world monitoring should focus on identifying blind spots, such as inappropriate responses in rare but critical scenarios like suicidal ideation [11]. Feedback loops—leveraging detailed human evaluations and empathy assessment frameworks such as EMRank—can incrementally improve safety and therapeutic value [100]. Additionally, frequent retraining using updated clinical practices and more representative, diverse datasets ensures that these systems remain both ethically aligned and clinically effective [10].

In conclusion, overcoming the limitations and ethical challenges of incorporating LLMs into mental health care requires a holistic and multi-pronged approach. Privacy safeguards, bias mitigation, explainability, accountability, and iterative refinement form the foundation of responsible LLM deployment. By harnessing interdisciplinary collaborations and aligning LLM capabilities with clinical and ethical standards, these systems can reliably augment mental health services. Building on the technological strides outlined in earlier discussions, future research should emphasize multimodal integrations and real-time safety audits, ensuring that LLM-driven tools become equitable, transparent, and trustworthy allies in advancing mental healthcare.

5 Personalization and Accessibility in Mental Health Interventions

5.1 Adaptive Conversational Frameworks

Adaptive conversational frameworks in the context of large language models (LLMs) represent a transformative approach to mental health interventions by tailoring interactions to users' individual needs, language preferences, and emotional states. This personalization enables LLMs to provide more responsive and supportive conversational experiences, enhancing user satisfaction and fostering therapeutic alignment. Central to these frameworks is the dynamic adaptation of dialogue systems, which relies on real-time data assimilation and nuanced understanding of user-specific contexts.

At the core of adaptive conversational frameworks are advancements in sentiment analysis and emotional intelligence, where LLMs, such as GPT-4 and fine-tuned variants, identify emotional cues from user inputs. By leveraging sentiment detection techniques to classify emotions like sadness, fear, or frustration, LLMs can modulate their tone and provide contextual responses [5]. These sentiment-driven interactions are critical for aligning conversational flow with the user's current mental state. For example,

strategies for dynamic tone adjustment, embedded in fine-tuned models like Psy-LLM, demonstrate effectiveness in tailoring therapeutic content while emphasizing emotional validation and support [10]. However, challenges persist in accurately identifying complex emotional states, especially when handling multi-layered user inputs with ambiguous emotional undertones.

Another significant mechanism in adaptive frameworks involves leveraging conversational history and metadata to generate contextually relevant responses. By retaining information from prior interactions—for instance, a user's mental health goals, expressed concerns, or linguistic style—LLMs can ensure continuity and coherence across dialogues. Models like MindGuide integrate conversational memory structures through tools such as buffer-based memory systems, enabling them to simulate human-level recall during prolonged interactions [101]. Although these frameworks enhance personalization, the reliance on historically stored data raises concerns regarding privacy and compliance with regulations such as GDPR, as noted in previous evaluations [3].

Goal-oriented conversation structuring is another hallmark of adaptive frameworks. By aligning responses with user-defined objectives, such as stress reduction, adherence to therapy strategies, or cognitive reframing, LLMs can contribute to meaningful therapeutic advancements. Approaches like RESORT, which employs principles of cognitive reappraisals, illustrate how LLMs guide users toward reevaluating negative appraisals of their circumstances by modeling structured and psychologically grounded responses [60]. This structured adaptability demonstrates high efficacy in mental health applications when combined with instructional prompts. However, studies reveal that misalignment with user-specific goals or poorly calibrated responses can inadvertently lead to user frustration or therapeutic disengagement [9].

Despite these advancements, limitations remain in adaptive personalization. One pressing challenge is the presence of biases in training data, which can skew LLM interpretations of user inputs, particularly for underrepresented groups. For instance, discrepancies in empathy detection between demographic subgroups, as observed in LLM evaluations, highlight the need for fair and inclusive training methodologies [12]. Additionally, the propensity for LLMs to produce hallucinated outputs can compromise the reliability of adaptive frameworks in high-stakes scenarios, demanding rigorous safeguards and domain-specific tuning [15]. Addressing these issues will require a synergy of bias mitigation techniques, such as counterfactual data augmentation, and the development of more robust evaluation benchmarks for nuanced personalization [102].

Emerging trends, such as multimodal integration, offer avenues for further refinement of adaptive frameworks. By combining textual inputs with additional data streams like speech prosody, facial expressions, or physiological signals, models can capture richer, multimodal representations of user states. This has been demonstrated in studies where LLMs, when paired with EEG and audio data, improved their ability to assess depression and anxiety symptoms [13]. Additionally, the development of lightweight, resource-efficient models shows promise in bringing adaptive mental health support to underserved communities, particularly where computational resources are limited [5].

In conclusion, adaptive conversational frameworks enabled by LLMs hold immense potential in enhancing mental health interventions through personalized, goal-driven, and emotionally intelligent interactions. Nonetheless, addressing challenges related to bias, hallucination, contextual limitations, and ethical considerations is paramount for realizing their full potential. Future research should prioritize integrating explainable AI methods, refining multimodal capabilities, and establishing standardized frameworks for evaluating personalization effectiveness in diverse real-world contexts.

5.2 Multilingual and Cross-Cultural Capabilities

Large Language Models (LLMs) hold immense potential to enhance mental health interventions by addressing linguistic and cultural barriers, thus improving accessibility to care for diverse global populations. As mental health challenges transcend geographic and cultural boundaries, effective interventions require alignment with users' unique linguistic preferences, cultural contexts, and communication styles. LLMs, including GPT-4 and LLaMA variants, demonstrate significant promise in meeting these needs through their advanced multilingual understanding, cultural sensitivity, and inclusive natural language processing (NLP) capabilities. However, technical, ethical, and resource-related challenges must be overcome to fully realize this potential.

Multilingual fluency stands out as a core strength of LLMs. Trained on vast, multilingual corpora, LLMs excel in supporting users across high-resource and several mid-resource languages by detecting intricate psychological states within multilingual contexts. Frameworks like RoBERTa and GPT-based models have shown effectiveness in identifying and classifying mental health indicators across linguistic borders [16], [21]. Innovations such as unsupervised machine translation and retraining with diverse, bilingual datasets have enhanced the multilingual capabilities of transformer-based models, addressing gaps in traditional NLP approaches [53]. Despite these advancements, significant gaps persist in the support of low-resource languages and dialects, which are often excluded from pretraining corpora due to limited availability of annotated data [14]. Addressing these limitations is crucial for fostering truly inclusive mental health systems.

Cultural alignment is equally vital in mental health discourse, as cultural contexts shape emotional expressions, symptom manifestations, and coping mechanisms. Models that fail to incorporate cultural context risk misinterpreting user inputs or producing irrelevant responses, potentially exacerbating mental health stressors. For example, the expression of depressive symptoms or psychological stress varies significantly across cultures, as evidenced by different linguistic markers and coping strategies among diverse populations [96]. LLMs have been fine-tuned on datasets tailored to culturally specific patterns of language and behavior. Cognitive Behavioral Therapy (CBT)-oriented models that infuse responses with cultural relevance have yielded higher user satisfaction [65]. However, achieving true cross-cultural congruence remains a complex challenge,

with models needing to balance universal applicability with nuanced cultural details.

Bias in training data is a pervasive obstacle impeding the cross-cultural adaptability of LLMs. Societal biases embedded in pretraining datasets often lead to disparities in responsiveness and empathy for underserved groups, disproportionately affecting marginalized users. For instance, lower empathy levels in responses to Black users compared to other demographics underscore the systemic biases that persist in existing models [12]. Strategies like counterfactual data augmentation, fairness-driven pretraining methods, and reinforcement learning from human preferences (RLHF) using demographic fairness metrics have been proposed to mitigate these biases. While these approaches show promise, their success requires significant investments in high-quality, diverse annotations and rigorous evaluation protocols [11].

Supporting underrepresented languages and dialects remains a high priority for increasing global accessibility. Advanced multilingual resources, including SMHD and curated datasets for underrepresented languages, have improved LLMs' ability to process cross-linguistic contexts [14], [30]. Recently, cross-lingual transfer learning and few-shot capabilities have allowed models to generalize to previously unsupported languages, offering scalable applications in mental health interventions [14]. However, practical deployment in low-resource regions is often hindered by technological barriers like high computational demands and inconsistent internet access. Lightweight architectures tailored for low-compute environments or offline-capable models addressing privacy and connectivity constraints hold significant promise for overcoming these challenges [13].

Future efforts should focus on enriching the linguistic and cultural diversity of training datasets, ensuring that overlooked languages and marginalized communities are prioritized in model development. Interdisciplinary collaboration between NLP researchers, mental health experts, and sociolinguists is essential for refining models to capture cultural nuances accurately. Moreover, ethical design principles, transparency, and user-centered frameworks must guide the creation of privacy-conscious, culturally sensitive LLMs for mental health care. While substantial progress has been made in reducing linguistic and cultural barriers, intensifying efforts to address biases, expand language coverage, and develop resource-efficient architectures will determine LLMs' potential to democratize mental health care on a global scale.

5.3 Accessibility in Resource-Limited Settings

Accessibility to mental health care in resource-limited settings represents a pressing global challenge, exacerbated by shortages of trained clinicians, infrastructural limitations, and pervasive stigma surrounding mental health. Large language models (LLMs) hold transformative potential in bridging these gaps by delivering scalable, cost-effective, and context-sensitive support. This subsection delves into the specific mechanisms by which LLMs enhance accessibility, their proven and theoretical capabilities, and the challenges and trade-offs inherent in deploying such systems in underserved regions.

Firstly, the modularity and adaptability of LLMs enable deployment in regions with limited technical infrastructure. Lightweight LLM architectures designed for efficiency, such as parameter-reduced transformer models, are critical for operating in low-compute devices and environments where intermittent connectivity is a barrier [71]. Serving as the backbone of offline solutions, these models can leverage advances in quantization techniques and compression algorithms, ensuring they remain functional without compromising significantly on core performance. For example, fine-tuned models such as Health-Alpaca have demonstrated efficacy in health prediction tasks while significantly reducing computational costs, making them apt candidates for integrations into local healthcare ecosystems [71]. The development of low-powered models would allow for wider adoption in regions without consistent access to reliable internet or state-of-the-art hardware.

Offline-first language models further enhance accessibility by addressing significant privacy concerns that often inhibit the adoption of digital interventions for mental health. Locally deployable systems reduce reliance on external servers, ensuring sensitive mental health data remains secure and aligned with users' rights to confidentiality. Zero-shot and few-shot learning paradigms, which have shown competitive performance in mental health classifications such as suicidality and depression detection [15], present an opportunity for implementing relatively untrained models that can still be adapted easily to cultural and linguistic nuances without requiring expansive local datasets. Models such as GPT-4 and Alpaca-LoRA have demonstrated robustness in handling sparse data environments, a necessity for resource-constrained regions where labeled mental health datasets are scarce [5].

The multilingual capabilities of LLMs could act as a cornerstone in addressing the linguistic diversity in these settings. Models can be fine-tuned to handle underrepresented languages and dialects, expanding their applicability to rural areas that rely on indigenous forms of communication. Papers such as "SMHD: A Large-Scale Resource for Exploring Online Language Usage for Multiple Mental Health Conditions" emphasize the role of fine-tuned datasets in catering to diverse linguistic representations of mental health symptoms. Culturally adapted conversational prompts, coupled with reinforcement learning from human preferences (RLHF), present a structured pathway for models to better capture regional expressions of distress or mental health symptoms, which vary widely in presentation due to cultural factors [5].

Despite their immense potential, deploying LLMs in resource-limited contexts involves significant trade-offs and challenges. For instance, the generalization capabilities of models often falter when exposed to culturally or linguistically unfamiliar datasets, introducing biases that can perpetuate inequitable outcomes. Research on demographic bias suggests that algorithms trained on predominantly Western data fail to account for the subtle linguistic markers of mental health conditions in non-English or low-resource languages [35], [36]. Moreover, the ethical risks of over-reliance on automated systems for mental health care cannot be understated. Ensuring transparent communication that LLMs serve as supportive tools—not replacements for hu-

man clinicians—remains critical in promoting appropriate usage and preserving local trust in interventions [11].

Leveraging community-based validation systems is crucial for ensuring model reliability and contextual alignment. Engaging local communities in co-design processes, wherein clinicians, patients, and developers collaborate to annotate data and adapt LLM outputs, offers a feasible way to improve model effectiveness while embedding culturally sensitive practices. Additionally, ongoing integration with multimodal systems promises to elevate the efficacy of LLMs in constrained environments through better interpretation and analysis of combined textual, audio, and visual data streams [40].

Finally, the future of increasing accessibility through LLMs in resource-limited regions hinges on interdisciplinary collaboration. Governments, NGOs, and private technology organizations should invest in creating open-source, low-cost models tailored for underserved populations, such as those demonstrated by "MentaLLaMA: Interpretable Mental Health Analysis on Social Media with Large Language Models," an open-source LLM optimized for interpretable mental health contexts. Iterative training processes, adaptive scaling, and engagement with ethical frameworks can position LLMs as scalable, inclusive instruments for democratizing mental health care worldwide.

By addressing computational, cultural, and infrastructural barriers simultaneously, LLMs exhibit profound potential to reduce disparities in global mental health services. Nevertheless, sustainable pathways require long-term commitments to addressing known limitations while fostering innovations that integrate local knowledge and technological advancement responsibly.

5.4 Ethical and Privacy Considerations in Personalization

The personalization of mental health interventions using large language models (LLMs) offers transformative potential to deliver more effective and user-centered support by tailoring interactions to an individual's unique context, preferences, and needs. However, it also presents significant ethical and privacy challenges that must be carefully navigated to avoid undermining the trust, equity, and safety required for these applications to succeed. This subsection critically examines these challenges, with particular focus on risks such as data dependency, systemic bias, user autonomy, and the safeguarding of confidentiality.

Personalization hinges on the collection and analysis of vast amounts of sensitive user data, including conversational histories, emotional states, and demographic attributes. While this granularity enhances the effectiveness of interventions, it also raises serious privacy concerns. Unauthorized access, data leaks, or misuse of user information for purposes beyond mental health care represent significant risks. Differential privacy techniques—where noise is injected into the data to protect individual identities—hold promise for mitigating some of these concerns, but their effectiveness depends on carefully balancing privacy preservation with maintaining personalization accuracy [14], [68]. Additionally, the fine-tuning of models on smaller, domain-specific datasets can minimize reliance on larger data repositories but still risks revealing sensitive patterns if anonymization procedures are insufficiently robust [14].

Systemic bias in personalized interactions poses another significant ethical challenge. Because personalization relies on algorithmic decision-making derived from training data, any embedded biases related to race, gender, age, or cultural background can unintentionally propagate inequities. Research has revealed that LLMs frequently generate differential outputs for stigmatized demographics, disproportionately associating mental health conditions with certain groups [48], [77]. These biases not only perpetuate health disparities but can reinforce harmful stereotypes, particularly for underrepresented populations. Strategies like fairness-aware prompting and counterfactual data augmentation show promise in mitigating such issues. However, their implementation must be both systematic and nuanced to ensure fairness without diminishing the fidelity of personalized interventions [78].

Another critical concern is the risk of fostering dependency on automated systems, potentially eroding user autonomy. Over-reliance on LLMs for emotional support or cognitive reframing could shift users towards treating these models as primary sources of assistance rather than complementary tools to human clinicians. This dynamic is particularly problematic when personalization amplifies the perception of deep emotional understanding, leading users to entrust sensitive issues to the system rather than seeking professional care [67], [74]. Designers must embed safeguards that encourage users to seek human intervention during high-stakes scenarios, especially when signs of severe distress are detected [11]. This safeguard-oriented approach will not only promote autonomy but also mitigate risks associated with misjudgment or system limitations.

Transparency is fundamental for fostering trust in personalized LLMs. Users must clearly understand how their data is utilized, the extent of personalization, and the boundaries of system functionality. Explainable AI methodologies have emerged as a crucial tool in this domain, enabling models to articulate the rationales behind their outputs. This transparency fosters accountability and reassures users about the safety and integrity of interactions [75], [103]. Furthermore, interdisciplinary collaboration involving technologists, clinicians, and ethicists is necessary to establish ethical oversight mechanisms that align personalization strategies with clinical and moral standards.

Looking ahead, addressing the technical trade-offs between personalization and privacy will be pivotal. Emerging approaches such as federated learning allow models to analyze decentralized datasets, minimizing risks associated with centralized data storage while still supporting effective personalization [94]. Concurrently, refining models through iterative feedback loops and audits is essential to reduce bias and ensure demographic fairness, with structured benchmarks providing a reliable framework for evaluation [42]. As mental health applications increasingly integrate LLMs, embedding ethical safeguards into the foundational design of these technologies will be critical to sustaining therapeutic value while upholding user rights and dignity.

5.5 Enhancing Therapeutic Relationships through Personalization

The advent of large language models (LLMs) in mental health applications offers a transformative opportunity to enhance therapeutic relationships through advanced personalization. Personalization in this context refers to the capacity to tailor conversational outputs to the unique emotional, cultural, and psychological needs of individuals, which is critical for fostering trust and connection in digital interventions. This subsection examines the potential of LLMs to simulate human-like empathy, build trust in automated interactions, and maintain therapeutic boundaries—all while addressing the associated technical and ethical challenges.

LLMs, such as ChatGPT and fine-tuned variants like MentalBERT, exhibit an impressive capability to process vast amounts of contextual data and generate responses that align with user-specific dynamics. Through personalization, these models adapt their linguistic style, tone, and content based on user input, enabling interventions that resonate more deeply with users' lived experiences [10], [15]. For example, contextual understanding is enhanced by integrating prior interactions and metadata, which LLMs utilize to craft responses imbued with historical coherence. This ability to scaffold conversations over time mirrors elements of long-term therapeutic alliances that are hallmarks of high-quality counseling. Key works [27] demonstrate how these features can be operationalized, particularly in interventions emulating cognitive-behavioral therapy (CBT) techniques.

However, fostering authentic empathy in LLMs remains both a promise and a challenge. Models achieve empathic responses through fine-tuning on datasets that contain labeled emotional and social cues while leveraging embeddings designed to track sentiment and emotional shifts in real-time. Studies have shown that incorporating domain-specific data, such as conversational corpora from psychotherapy sessions, enhances the perceived emotional acuity of responses [104]. Despite these advancements, the simulation of empathy by LLMs is fundamentally mechanistic, raising concerns about their ability to consistently navigate the subtle intricacies of human emotions. Misaligned responses or "hallucinations"—outputs that deviate from user needs or therapeutic norms—can impair trust and, in some cases, exacerbate psychological distress [29], [52].

Trust-building is an essential component of therapeutic relationships, and automation introduces unique dynamics in this regard. Transparency about LLM functionality, limitations, and ethical guidelines enhances user trust by aligning expectations with reality [29]. For instance, clearly communicating that the system serves as an adjunct rather than a replacement for professional care helps frame the appropriate application of these tools, safeguarding against over-reliance. Research further suggests that coupling accurate diagnostic insights with personalized psychoeducation fosters user empowerment, a quality strongly correlated with positive therapeutic outcomes [66].

While LLMs offer unprecedented scalability, maintaining therapeutic boundaries remains critical to ensure professional integrity. The conversational adaptability of LLMs, though advantageous, carries risks of overstepping clinical directives or projecting a false sense of intimacy, which may blur the boundaries between human competencies and artificial ones [83], [105]. Prompt engineering and reinforcement learning strategies are being investigated as potential mitigations to guide LLM outputs within ethical and therapeutic constraints [106].

Emerging integration of multimodal data, such as audio and visual cues, into LLM frameworks represents a promising frontier in personalization. Integrating speech prosody, facial expressions, and linguistic analysis allows for richer contextual awareness, mirroring the nuanced observations of human therapists [17], [31]. Such advancements suggest that future implementations could simulate dyadic therapist-client interactions more faithfully, with potential applications in complex therapeutic scenarios.

Looking ahead, enhancing therapeutic personalization with LLMs demands interdisciplinary collaboration between computational scientists, psychologists, and clinicians. Rigorous clinical validation, ethical adherence, and user trust will remain pivotal to their successful adoption. By addressing these challenges thoughtfully, LLMs hold the potential to revolutionize the depth and breadth of mental healthcare delivery, particularly in underserved communities where scalable, personalized interventions are most needed.

6 Ethical, Safety, and Societal Challenges in Mental Health Applications

6.1 Privacy and Data Security Concerns

Modern applications of large language models (LLMs) in mental health services introduce significant privacy and data security challenges due to the intimate and sensitive nature of mental health data. These challenges emerge from the intrinsic characteristics of LLMs that involve the collection, storage, and processing of large volumes of user interactions, often containing highly personal information. Ensuring the confidentiality of such data while addressing the technical and regulatory hurdles surrounding its use is paramount to fostering user trust and advancing ethically sound applications.

A primary concern lies in the proper handling of sensitive mental health information within LLMs, such as symptoms, diagnoses, emotional vulnerabilities, and personal narratives submitted during interactions. Unlike traditional mental health tools, which may operate within established boundaries of trust between patient and clinician, LLM-based systems typically rely on extensive backend infrastructures that include data transfer, cloud storage, and third-party integrations. Such systems create multiple points of vulnerability where breaches, unauthorized access, or data misuse could occur. Studies highlight the susceptibility of AI systems to model inversion or re-identification attacks, wherein adversaries can infer sensitive user information from aggregated training data [7], [107]. Re-identification is especially problematic in mental health because anonymized data, such as text excerpts describing traumatic experiences, can often reveal identifiable information through subtle linguistic cues—risks significantly amplified by the contextual sensitivity of mental health content.

Effective anonymization methods, such as differential privacy, have shown promise in mitigating these risks while maintaining data utility for model training. Differential privacy introduces statistical "noise" to datasets to obscure individual contributions, thereby safeguarding user identities [63], [85]. However, the trade-off between maintaining privacy and preserving model performance remains a formidable barrier. Contextual language modeling often requires high fidelity to nuanced linguistic expressions, something that can be compromised by aggressive anonymization techniques. In mental health services, loss of these details could impair diagnostic accuracy or the generation of appropriate therapeutic responses, creating a tension between privacy assurances and functional efficacy.

Legal compliance presents another crucial facet of navigating data security. Mental health applications leveraging LLMs must adhere rigorously to frameworks like the General Data Protection Regulation (GDPR) in Europe and the Health Insurance Portability and Accountability Act (HIPAA) in the United States. These regulations require robust mechanisms for securing personal data, explicit definitions of consent, and clear boundaries for data use. However, challenges arise from the global deployment of many LLM-based systems, where jurisdictional overlaps and conflicts in privacy regulations complicate compliance efforts. Bridging these regulatory gaps requires multistakeholder collaboration, with some researchers proposing modular frameworks that employ adaptive privacy-preserving techniques based on regional requirements [1].

Beyond direct data security, the potential for misuse or unauthorized secondary data usage poses ethical dilemmas. Compounding this issue is the opaque nature of many LLMs, which rely on vast pretraining datasets sourced from the web or proprietary databases without explicit disclosure of data lineage or processing mechanisms. Such opacity conflicts with principles of transparency and accountability, essential for ensuring ethical deployment in mental health contexts [108]. Efforts to enhance transparency are emerging, such as the introduction of audit tools to trace both individual and aggregate contributions to pretraining data and the wider adoption of open-source medical LLM frameworks, such as Hippocrates, which provide full access to training and evaluation processes [108].

To mitigate risks further, real-time solutions such as federated learning—where data remains on local devices and only aggregated model updates are transmitted—show promising applicability. Applicability studies have demonstrated that federated learning reduces the risk of exposing sensitive user interactions while maintaining adaptable model performance over time [63]. Additionally, secure multi-party computation (SMPC) provides opportunities for collaborative model development across institutions without directly sharing sensitive data, which could benefit global mental health initiatives requiring cross-institutional cooperation [7].

Emerging trends in privacy-aware personalization also hold potential to balance user expectations of secure and relevant interactions. Techniques like zero-shot personalization or content masking—where users can selectively control which portions of their interactions are retained for further processing—are being integrated into conversational agents to help build trust in LLM-powered tools [109].

The paramount task for researchers, clinicians, and technologists is to strike a thoughtful balance between advancing model capabilities and addressing ethical imperatives. Future directions must prioritize iterative collaboration among computer scientists, ethicists, and mental health professionals to design privacy architectures tailored to the unique sensitivities of mental health applications. Concurrently, rigorous audits, robust legal alignments, and real-world validations should become integral to the lifecycle of LLM-based mental health tools. Only with holistic, privacy-centric approaches can the promise of LLMs for mental health services be fully realized while safeguarding the dignity and rights of individuals.

6.2 Risk of Misinformation and Harmful Outputs

The deployment of large language models (LLMs) in mental health applications introduces significant risks of misinformation and harmful outputs, posing dire consequences for individuals relying on these tools during vulnerable moments. These concerns emerge from the interplay between the inherent limitations of LLMs, their operational contexts, and the specific demands of mental health care. This subsection examines the origins of these risks, their practical implications, and strategies for effective mitigation, creating a bridge between the ethical concerns around data privacy discussed earlier and the critical issues of bias and fairness that follow.

LLMs, by design, are probabilistic systems trained on extensive, often uncurated datasets, which makes them susceptible to generating "hallucinations"—fabricated or incorrect information presented as factual. In the context of mental health, such hallucinations can lead to diagnostic inaccuracies or the dissemination of unsafe advice in critical situations. For instance, an LLM tasked with identifying symptoms or offering support could erroneously flag linguistic markers of mental health issues, resulting in false positives or overlooking urgent risks like suicidal ideation [22], [69]. This misalignment between general-purpose language modeling and the specificity required for clinical-grade interventions further underscores the challenges of using LLMs in sensitive mental health applications [28].

A critical factor driving these risks is the lack of domain-specific alignment in many LLMs. Most models are pretrained on generalized corpora and thus lack the nuanced understanding essential for evidence-based mental health care. Without fine-tuning on rigorously validated, domain-specific datasets, these models often fail to align with clinical guidelines, increasing the risk of generating misleading recommendations. For example, LLM-powered systems may inaccurately distinguish between disorders with overlapping linguistic markers, such as anxiety and depression, potentially leading to inappropriate interventions or inaction in critical cases [16], [19].

Similarly, a tendency to generate overly broad, generic, or surface-level responses exacerbates the issue, particularly in sensitive psychological interactions. When users seek personalized support, LLMs might offer advice grounded in shallow textual patterns, resulting in overly simplistic or dismissive outputs [67]. For example, optimistic responses

intended to console users may inadvertently invalidate their emotional experiences, causing further emotional distress and harming therapeutic outcomes [64].

Beyond the technical challenges, the societal implications of misinformation in mental health applications cannot be neglected. Incidents of LLM-powered systems producing inappropriate or harmful advice highlight the potential erosion of trust in digital health tools. This reputational risk underscores the need for comprehensive safeguards and robust quality assessments to ensure that such applications meet the critical demands of mental health care [12], [15]. Without systemic checks and feedback loops to quickly identify and correct inaccuracies in high-stakes contexts, LLMs cannot reliably function as mental health support tools.

To mitigate these risks, embedding robust operational guardrails into LLMs is essential. Techniques such as fine-tuning models with domain-specific data and leveraging instruction-based learning frameworks can improve the alignment of LLMs with clinical practices [5], [24]. Real-time oversight systems, such as human-in-the-loop configurations, allow mental health professionals to contextualize and verify outputs, minimizing the potential harm caused by misinformation [68]. Additionally, explainability frameworks play a critical role in enabling clinicians to interrogate and assess the reliability of model-generated responses, fostering greater transparency and safer applications [28], [69].

Looking ahead, the iterative evaluation and benchmarking of LLMs for safety in mental health contexts must become a priority. Metrics such as hallucination rates, conversational empathy, and alignment with diagnostic protocols should be systematically incorporated into performance assessments [68]. Hybrid approaches that integrate multimodal inputs, such as speech and facial cues, hold promise for enhancing contextual understanding and reducing misinformation in mental health interventions [17], [31].

In sum, while LLMs have immense potential to democratize access to mental health resources, their deployment must be approached with caution given their susceptibility to misinformation and harmful outputs. Addressing these limitations requires technical, clinical, and regulatory collaboration to align these systems with the rigorous standards of mental health care. By embedding safeguards, fostering transparency, and adhering to ethical principles, LLMs can evolve into reliable, supportive tools, complementing human-driven mental health interventions while minimizing risks. This commitment to balancing innovation with safety will ensure LLMs fulfill their promise of equitable and effective mental health support.

6.3 Bias and Fairness in Mental Health Applications

Bias and fairness in large language models (LLMs) deployed for mental health applications represent a pressing concern, as such systems, when inadequately scrutinized, can perpetuate and even amplify systemic inequities. Mental health care already suffers from disparities, disproportionately affecting marginalized groups based on race, gender, socioeconomic status, and geographic location. The deployment of LLMs trained on biased or incomplete data risks exacerbating these inequities, leading to reduced quality of care for vulnerable populations.

At the core of this issue lies the quality and representativeness of training data. Many LLMs are trained on datasets predominantly sourced from specific linguistic, cultural, and social contexts, often privileging Western, English-speaking, and middle-to-upper-class users [69], [72]. Consequently, these models may fail to generalize adequately for diverse populations. For example, cultural expressions of distress or linguistic markers of mental health conditions can differ significantly, leading to systematic under-diagnoses or over-diagnoses of mental health issues in certain groups [5]. Similarly, underrepresented languages and dialects complicate equity further, as even state-of-the-art models like GPT-4 struggle to achieve comparable accuracy across non-dominant linguistic contexts [35], [72].

Demographic imbalances in model training data also give rise to skewed outputs. Studies have shown that LLMs can reflect and perpetuate harmful stereotypes—such as associating certain personality traits or mental health conditions with specific racial or gender groups—due to the latent biases in their datasets [1], [72]. For example, mental health conditions like depression or anxiety might be over-predicted among marginalized groups due to linguistic patterns improperly flagged as risk markers during training [18]. However, algorithms may simultaneously underpredict distress manifestations in populations whose psychosocial and linguistic markers diverge from the majority group, leading to a failure to intervene for those who need care the most [36].

Intersectionality compounds these challenges, as layered identity factors significantly impact mental health outcomes and access to care. For instance, low-income queer youth of color may face intersecting biases in both the data used to train LLMs and the model's failure to fully account for complex identity-specific mental health needs [29], [73]. Addressing such compounded disparities requires an evaluative framework that goes beyond single-axis fairness metrics and critically examines intersectional representation within datasets and across predictive outcomes.

Efforts to mitigate bias in mental health applications of LLMs have explored several technological strategies. Counterfactual data augmentation, which involves purposefully introducing synthetic examples of minority-centric use cases, has been shown to reduce demographic disparities in output predictions by increasing model exposure to underrepresented narratives [5]. Additionally, fine-tuning models with fairness-aware datasets (e.g., datasets curated to balance demographic attributes such as gender and racial representation) has improved their performance for diverse user groups [30]. However, such methods are not without trade-offs—bias mitigation techniques often improve fairness at the cost of overall model efficiency and accuracy for the majority population [110].

Evaluation frameworks for mental health-focused LLMs further complicate bias detection and reduction. The absence of standardized benchmarks tailored to mental health disparities makes it difficult to systematically assess fairness. Additionally, widely accepted datasets, like PHQ-9 or DSM-aligned corpora, often fail to incorporate culturally sensitive indicators of mental illness, further perpetuating

exclusion [69], [111]. Methodological tools like saliency maps, which seek to identify which input features most influence LLM outputs, can serve as a partial solution, offering transparency into how demographic markers influence predictions [42]. However, scalability and implementation ease remain unresolved issues.

Looking ahead, the field must move beyond surface-level bias mitigation, integrating robust cross-disciplinary collaborations among technologists, clinicians, and ethicists to develop equitable systems. Fairness-aware pretraining pipelines, continuous dataset evaluation, and intersectional performance metrics are critical steps for advancing this aim. Moreover, involving representatives from marginalized communities in system design can ensure that the unique needs of these populations are explicitly accounted for in all phases of model development and use [29].

By systematically addressing bias at both the data and algorithmic levels, the deployment of LLMs for mental health interventions holds immense transformative promise. However, achieving this potential requires sustained commitment to fairness, rigorous empirical validation, and a willingness to prioritize ethical considerations alongside algorithmic performance.

6.4 Transparency and Explainability of LLM Outputs

As the adoption of large language models (LLMs) expands in mental health applications, addressing transparency and explainability has emerged as a critical imperative. These dimensions are particularly essential in clinical contexts, where patients and clinicians depend on LLM outputs for sensitive, high-stakes mental health decisions. Without clear insights into how decisions are derived, the risks of misaligned, biased, or harmful outputs remain unacceptably high. Building on the discussion of bias and fairness in the previous section, this subsection explores the technical, practical, and ethical challenges associated with ensuring transparency and interpretability in LLM-driven mental health tools, highlighting emerging methods and identifying key areas for improvement.

Transparency refers to the ability of models to provide a comprehensible account of how their outputs are generated—a concept that becomes especially challenging with the inherent complexity of LLMs. These deep learning systems, such as GPT-4 and LLaMA-2, operate with vast numbers of parameters and layers, resulting in "black box" characteristics that make their decision-making processes opaque to human observers. In mental health, where clinicians must understand and validate the basis of a model's recommendation to ensure trust and accuracy, such opacity is a significant barrier [74], [75]. Techniques like attention visualization, saliency maps, and Layer-wise Relevance Propagation offer glimpses into the internal workings of LLMs by identifying which input features (e.g., specific words or phrases) contributed most to an output [103], [112]. Yet, these methods often provide only partial transparency and fail to fully elucidate the causal pathways shaping model outputs.

Explainability, on the other hand, extends beyond technical transparency to encompass the model's ability to convey its reasoning in a manner that is intelligible to non-expert users, such as patients or caregivers. This aspect is paramount in mental health contexts, where individuals may already be in states of emotional vulnerability. Promising advances include the generation of natural language explanations to accompany diagnostic predictions or therapeutic suggestions. For example, MentalLLaMA, an open-source model fine-tuned for mental health analysis, integrates interpretability through task-specific instruction fine-tuning to explain its predictions based on clinical evidence [38]. Similarly, integrating structured psychiatric scales, such as the Columbia Suicide Severity Rating Scale (C-SSRS), into LLM workflows helps produce clinically validated and more interpretable outputs [76], [111]. However, while such methods are notably promising, they face limitations in ensuring the coherence, specificity, and precision necessary for high-stakes mental health applications.

Challenges to achieving effective explainability are multifaceted, spanning both technical and practical dimensions. One pressing issue is the phenomenon of hallucination, in which LLMs generate factually inaccurate outputs with unjustified confidence, potentially misleading clinicians or exacerbating patient distress [75], [93]. Striking a balance between model complexity and interpretability poses another unresolved challenge. Simplifying LLMs to enhance their explainability often compromises their diagnostic precision, particularly when addressing nuanced and intersectional mental health conditions [91], [103]. Furthermore, mental health is a highly contextual domain, influenced by linguistic, cultural, and socio-demographic factors. As such, generalized LLMs may fall short in capturing the nuances required for both accurate predictions and tailored explanations.

Emerging solutions aim to tackle these challenges through innovative approaches to transparency. Post hoc explainability tools, for instance, enable secondary systems to dissect and contextualize LLM outputs, while proactive transparency mechanisms, such as multi-modal alignment, integrate textual analysis with complementary patient data sources, including voice tone and facial expressions [17]. Dynamic prompting mechanisms, such as dual-prompting designs, combine domain-specific reasoning with linguistic evidence extraction to refine explanations while safeguarding model performance [68], [112]. These techniques illustrate the potential for iterative and layered solutions to enhance both transparency and accuracy.

Looking ahead, addressing the persistent gaps in transparency and explainability requires collective efforts across technical innovation and interdisciplinary collaboration. Standardized frameworks for evaluating these dimensions specifically within mental health contexts are urgently needed, taking into account usability, clinical safety, and cultural responsiveness. Additionally, aligning LLMs with professional guidelines—such as those outlined in DSM-5 or ICD-11—offers a pathway for grounding explainable outputs in standardized clinical evidence [46], [113]. Real-time explainability mechanisms that foster dynamic, context-driven interactions between models and clinicians hold promise for increasing trust while mitigating risks [27].

Ultimately, advancing transparency and explainability in LLM-based mental health applications aligns closely with broader ethical imperatives, as explored in the following section. By ensuring that these models operate as inter-

pretable and accountable tools, their integration can complement human expertise in mental health care, fostering both safety and effectiveness in this sensitive domain.

6.5 Ethical Decision-Making and Responsibility in Deployment

The deployment of large language models (LLMs) in mental health applications raises multifaceted ethical challenges, requiring careful deliberation on decision-making frameworks and responsibility to optimize societal benefits while mitigating risks. This subsection explores ethical dilemmas associated with automated mental health support, the necessity of informed consent and user awareness, and the critical role of interdisciplinary collaboration in fostering ethical deployment practices.

A central ethical dilemma in deploying LLMs for mental health lies in balancing the potential to enhance accessibility and scalability with the risk of replacing or undermining human therapists. While leveraging LLMs can address systemic gaps in access to mental health services, there remains skepticism regarding whether these tools can replicate the nuanced understanding required for clinical care. As highlighted in [83], human clinicians bring unmatched empathy and nuanced interpretation, which LLMs cannot fully emulate. Over-reliance on LLMs for emotional and psychological guidance may limit opportunities for patients to seek professional care when necessary. This risk is particularly acute when LLMs assert advice that appears authoritative but diverges from evidence-based standards, as the "hallucination" trait of LLMs can lead to harmful outputs in high-stakes scenarios [29].

Furthermore, the opacity of LLM architectures introduces challenges concerning users' understanding of their capabilities and limitations. Ensuring informed consent requires equipping users with sufficient knowledge about how LLM-powered systems function, their potential biases, and how their data are used for interactions. In many cases, users may not fully grasp that interactions with LLMs are not equivalent to professional therapy, creating a false sense of security about the guidance provided [29]. This issue can be exacerbated in vulnerable populations, where individuals experiencing distress may unquestioningly rely on these systems for support. Ethical safeguards necessitate that developers clearly communicate system limitations and, where possible, provide transparency mechanisms such as explainability tools that promote trust without misrepresenting the system's capabilities [69].

Key to fostering ethical responsibility is embracing interdisciplinary collaboration that aligns the technical design of LLMs with clinical best practices and human-centered principles. As [106] emphasizes, incorporating domain-specific behavioral and interactional knowledge can enhance LLMs' alignment with therapeutic standards while also ensuring model decisions reflect rigorous psychiatric principles. However, constructing systems with such tailored designs raises pressing concerns about the trade-offs between generalizability and specialization. LLMs applied across diverse user contexts may fail to adapt appropriately to culturally specific norms or intersectional vulnerabilities, as discussed in [29].

The integration of contextual knowledge into LLMs must extend beyond coding processes to ethical governance structures. Interdisciplinary teams comprising technologists, clinicians, ethicists, and policymakers can collaboratively ensure mechanisms are in place to uphold user agency and adherence to established mental health standards [83]. Strategies such as establishing mandatory human oversight for critical interventions, regulatory auditing of algorithmic fairness, and appropriate triaging mechanisms for urgent cases can further mitigate risks. Enforcing these safeguards aligns with calls to treat LLMs as complementary tools to augment—not supplant—human expertise [29].

Looking ahead, advancing ethical decision-making frameworks will require iterative understanding of LLMs' long-term societal implications. Monitoring post-deployment effects, such as shifts in public attitudes toward traditional therapy or dependency on automated care, remains paramount. Integrating real-time user feedback, as explored in [29], can provide a responsive mechanism for ensuring LLM updates remain aligned with changing societal and clinical contexts. Ultimately, the responsible deployment of LLMs for mental health necessitates a balanced approach that reinforces human dignity while unlocking technologies' transformative capacities in addressing mental health challenges globally.

6.6 Ensuring Safety in Deployment and Long-Term Monitoring

Ensuring safety during the deployment and ongoing use of large language models (LLMs) in mental health applications is a critical challenge, given the sensitive nature of patient well-being and the potential for unintended harm. With ethical considerations forming the bedrock of mental health applications, safety protocols must address diverse risks—ranging from harmful outputs and hallucinations to biases in diagnosis or therapy delivery—while also adapting to evolving user needs and shifting paradigms in mental health care. Building on the ethical imperatives discussed earlier, this subsection explores pragmatic strategies to ensure safety, from pre-deployment testing to continuous monitoring and regulatory oversight, highlighting the interplay between technological innovation and human-centered design.

A cornerstone of safety assurance lies in rigorous pre-deployment testing, emphasizing controlled environments that simulate high-stakes real-world scenarios such as crisis intervention or suicide prevention. These simulations enable researchers and developers to identify failure points, especially in cases of complex, ambiguous, or emotionally sensitive inputs. Insights from studies [5], [68] illustrate the importance of designing benchmarks tailored specifically to mental health contexts. Unlike generic natural language processing (NLP) tasks, these benchmarks should assess nuanced dimensions such as empathy, coherence, and alignment with evidence-based therapeutic standards. Notably, recent research on evaluating empathy in LLMs [100], [114] offers valuable methodologies for measuring user-centered outcomes like emotional comfort and therapeutic validity. However, this approach often encounters a trade-off between comprehensive evaluation and computational

efficiency, as testing for such granular metrics can demand substantial resources.

Once deployed, maintaining safety over the long term requires dynamic, iterative monitoring processes. Real-time feedback mechanisms—wherein user input and clinician evaluations help refine system behavior—play a pivotal role in ensuring adaptability. For instance, reinforcement learning from human feedback has proven instrumental in calibrating conversational flow and appropriateness, as demonstrated by CaiTI [115]. Additionally, techniques such as prompt engineering refinement and dynamic rule-based filtering can mitigate risks of hallucinated or inappropriate outputs [12], [60]. However, distinguishing genuine feedback from malicious exploitation remains challenging, necessitating robust anomaly detection mechanisms that respect user privacy.

Regulatory and oversight frameworks provide an essential safety net, particularly when sensitive mental health data is at risk. Adhering to standards such as GDPR or HIPAA is imperative for ensuring user trust, though challenges emerge from the re-identifiability of conversational data given its linguistic nuances [13], [72]. Emerging privacy-preserving techniques, such as differential privacy and federated learning, offer promising solutions for safeguarding user confidentiality while enabling system improvements in real time. Embedding human experts into the operational workflow to review critical flagged interactions offers an additional layer of protection, reducing the likelihood of catastrophic failures [10].

The integration of multimodal data streams, including text, facial microexpressions, and neurophysiological signals like EEGs, introduces new dimensions to safety assurance mechanisms. These advances promise heightened diagnostic accuracy and improved early-warning systems [13], [17]. However, such innovations must be carefully calibrated to avoid compounding biases inherent to individual modalities, and to prevent potential conflicts across data streams.

Despite notable progress, significant challenges remain. Model drift due to incomplete training updates, the amplification of systemic biases, and the over-reliance on AI systems during moments of crisis highlight pressing vulnerabilities requiring attention. Suggestions for addressing these include implementing ongoing audits, incorporating domain-specific pretraining based on emerging psychological guidelines, and fostering close collaboration between technical experts and mental health practitioners [9], [97]. Moreover, transparent communication regarding the capabilities and limitations of deployed LLMs is vital for managing user expectations, maintaining trust, and reducing risks of potential harm.

In summary, the safety of LLMs in mental health applications demands a multifaceted approach that integrates robust technical innovations, ethical considerations, and proactive regulatory oversight. Bridging the insights from rigorous pre-deployment testing with strategies for long-term adaptation, these efforts must strive to harmonize rapidly advancing technologies with core principles of equitable, trustworthy, and empathetic care. As the landscape of mental health support evolves, the continued iteration and refinement of these systems will be necessary to maintain their reliability and responsibility, ensuring they complement and enhance traditional therapeutic practices rather than undermine them.

7 Real-World Deployments, Evaluation, and Benchmarks

7.1 Establishing Benchmarks for Large Language Models in Mental Health

The development of benchmarks to evaluate Large Language Models (LLMs) in mental health domains requires a multi-faceted and domain-specific approach, given the sensitivity and complexity of mental health applications. Traditional metrics in broader natural language processing (NLP) tasks, such as BLEU scores or general-purpose accuracy measures, are insufficient for evaluating the nuanced capabilities required for this context, such as empathy, therapeutic alignment, and cultural sensitivity. This subsection highlights the current efforts, challenges, and future directions in establishing suitable benchmarks that rigorously evaluate LLM-based mental health tools across diagnostic, therapeutic, and conversational applications.

One critical area of benchmark design centers around evaluating models' ability to simulate empathy and alignment with therapeutic principles. Metrics focused on emotional accuracy, empathy recognition, and conversational coherence are paramount in assessing LLMs embedded in mental health support systems [59]. Empathy, as a communicative mechanism, has multidimensional aspects, encompassing cognitive understanding of another's feelings, emotional resonance, and the ability to respond appropriately. Recent works propose frameworks such as multi-task fine-tuning models to rate empathy scores across interactions [59]. However, challenges remain in operationalizing empathy benchmarks, particularly for text-based mental health applications, where responses need to balance depth without veering into overly prescriptive or inappropriate advice.

Another dimension critical to benchmarking LLMs in this domain is diagnostic accuracy, which requires precise assessment of the models' ability to identify and differentiate mental health conditions like depression, post-traumatic stress disorder (PTSD), anxiety, or ADHD. Studies leveraging datasets like SMHD [14] or public social media data [4] evaluate LLMs on tasks such as binary classification of mental health conditions or multi-class prediction of comorbid disorders. Metrics such as precision, recall, specificity, and sensitivity provide insights into the diagnostic robustness of LLMs, while domain-specific refinements, including longitudinal assessments of sentence-level predictions, can enhance these benchmarks. While early studies demonstrate that targeted fine-tuning boosts diagnostic precision significantly [2], there is still a need to address racial, cultural, and gender biases emerging from training data [12].

Beyond diagnostic tasks, effective summarization of psychological data, such as therapy session notes or patient self-reports, represents an impactful benchmark due to its capacity to reduce clinician workloads. Studies employing medical LLMs, such as ClinicalGPT and GatorTron [57], [85], have addressed clinical summarization but remain largely unexplored in mental health contexts. Metrics for

summarization accuracy must go beyond syntactic completeness and concentrate on the preservation of patient narratives and affective nuances, ensuring that critical markers, such as mood shifts, are effectively captured.

Cultural and linguistic factors present unique challenges in benchmark diversification. Multilingual and culturally adaptive evaluation strategies are necessary to ensure LLM accessibility for global populations. Current benchmarks often fail to reflect dialectal nuances or culturally grounded expressions of distress, highlighting the need for datasets that represent under-resourced languages and communities. Techniques like counterfactual data augmentation and domain-adaptive pretraining [86] offer promising avenues to address these disparities.

Emerging trends show increasing interest in dynamic and multimodal benchmarks that integrate non-textual data along with LLM outputs, such as audio cues, facial expressions, or electroencephalogram (EEG) signals [13]. These benchmarks can capture more complex indicators of mental health conditions, improving the ecological validity of LLM evaluations. For instance, multimodal models incorporating language and EEG data outperform single-modality counterparts in depression detection, signaling future pathways for robust benchmarks in clinical research.

Finally, usability and safety remain cornerstone considerations for benchmark frameworks. User satisfaction and therapeutic outcomes—whether measured through controlled patient interactions or longitudinal studies—have begun to surface as integral metrics, accompanied by safety metrics to evaluate risks of psychological harm or incorrect outputs [74]. Future benchmarks should also prioritize evaluating real-time deployment scenarios, assessing how effectively models engage users repeatedly while avoiding pitfalls such as overfitting to past interactions or undermining professional therapeutic boundaries.

To advance the field, interdisciplinary collaborations between technologists, clinicians, and researchers are essential for designing benchmarks reflective of real-world requirements. Benchmark consortia could unify data standards, evaluation protocols, and ethical guidelines, ensuring scalable, equitable progress in mental health applications of LLMs [11]. Open datasets and tools, such as those proposed in frameworks like PMC-LLaMA [116], can accelerate these efforts while fostering transparency and accountability.

In summary, establishing domain-specific benchmarks for mental health-focused LLMs requires balancing technical rigor, clinical relevance, and ethical considerations. Emerging advancements in multimodal evaluations, empathy metrics, and fairness-aware methodologies underline this crucial endeavor. The future lies in crafting benchmarks that not only measure performance but also guide the safe, equitable, and impactful integration of LLMs into mental health care.

7.2 Evaluation of Real-World Usability and User Engagement

The evaluation of real-world usability and user engagement for large language model (LLM)-based mental health applications plays a crucial role in understanding their effectiveness, practicality, and long-term impact. Building on discussions of benchmarking and ethical considerations in previous sections, this subsection explores key methodologies and metrics employed to assess these applications' capacity to provide meaningful, sustained mental health support while identifying associated challenges and opportunities for refinement.

Usability in LLM-driven mental health tools is deeply rooted in user-centered evaluations that encompass accessibility, acceptability, and adaptability. Metrics such as user satisfaction—often measured through post-interaction surveys, self-reported ease of use, and subjective well-being improvements—offer valuable insights into LLM performance. For example, a study utilizing ChatGPT-based classification models for stress and depression detection highlighted significant user benefits, with an F1 score of 0.86 for depression detection, underscoring its potential for real-world deployment [15]. However, engagement metrics like retention rates, interaction frequency, and perceived relevance extend beyond satisfaction, underscoring the necessity of dynamic interaction strategies and tailored approaches to sustain user interest over time.

Personalization emerges as a pivotal factor influencing user engagement. Applications like Replika, which have been tested for mental well-being support, revealed that while users valued its on-demand availability, issues such as inconsistent communication and lack of memory about prior interactions posed barriers to sustained engagement [67]. A stronger focus on long-term engagement is demonstrated in adaptive learning frameworks, which refine responses based on prior user interactions to better align with users' evolving mental health needs. For instance, tailored interventions enabling cognitive restructuring achieved a 65

Despite this potential, challenges in maintaining consistent user engagement reveal risks such as over-dependence or counterproductive reliance on LLM-based tools. Metrics monitoring inappropriate usage patterns can mitigate some of these risks, as evidenced by the limited success of automated classifications in addressing nuanced user needs [83]. Moreover, demographic disparities in interaction outcomes, such as reduced empathetic responses for underrepresented populations, have raised concerns regarding inclusivity and fairness in LLM-based interventions [12]. These findings call for deeper work to enhance demographic generalizability and mitigate biases in engagement outcomes.

Practical engagement also entails addressing ethical and safety implications. Issues such as mitigating harmful or inappropriate outputs are critical to maintaining user trust. Advanced mechanisms, such as agentic frameworks that access real-time validated external knowledge, have significantly improved output reliability [68]. Furthermore, interpretability of LLM outputs remains essential in fostering user confidence. Systems that provide diagnostic explanations alongside predictions were perceived as more trustworthy and actionable in real-world mental health contexts [27].

Emerging trends in feedback loops and continuous model refinement offer promising directions for improving usability and engagement. For example, fine-tuned mental health systems such as Mental-FLAN-T5 and Mental-Alpaca, driven by instruction fine-tuning and task-specific adaptations, have outperformed baseline models in bench-

mark performance while enabling more nuanced mental health dialogue with users [5]. These developments highlight how real-time user data can refine interaction relevance and align LLM outputs with clinical standards.

In conclusion, while LLM-based tools introduce innovative, scalable, and personalized opportunities for mental health interventions, their real-world usability and engagement face several ongoing challenges. Enhancing demographic inclusivity, maintaining ethical safeguards, and advancing personalization capabilities will be crucial for fostering robust user retention and trust. As outlined by case studies in subsequent sections, prioritizing rigorous domain-specific fine-tuning and longitudinal evaluations will ensure these models adapt to evolving user needs while preserving equity, safety, and effectiveness. This interdisciplinary approach will drive the safe and impactful integration of LLMs into real-world mental health services.

7.3 Case Studies of Successful Deployments

The real-world deployment of large language models (LLMs) in mental health services underscores their tangible impact, providing critical insights into their potential to address global mental health challenges. This subsection examines select case studies where LLM-enabled solutions have been implemented to improve mental health service delivery, analyzing their successes, limitations, and the broader implications for the field.

One compelling example is the application of LLMs in social media-based early detection systems for mental health disorders. Models have been fine-tuned to identify psychological distress markers in user-generated content, such as posts containing linguistic patterns indicative of depression, anxiety, or suicidal ideation. For instance, a system leveraging GPT variants successfully demonstrated the ability to detect major depressive episodes and suicidal ideation from Reddit posts, achieving high precision through tailored prompts and domain-specific fine-tuning [35], [111]. Notably, the nuanced integration of clinical frameworks such as the Columbia Suicide Severity Rating Scale significantly enhanced the system's predictive accuracy during temporal risk assessments, illustrating how LLMs can bridge gaps between traditional clinical tools and digital interventions [76].

Another prominent implementation is the use of LLM-powered conversational agents in virtual mental health counseling. Applications like Replika and bespoke chatbots have successfully translated advanced therapeutic techniques, including motivational interviewing and cognitive restructuring, into digital interfaces. These solutions, integrated with pre-trained LLMs, exhibited robust capabilities in maintaining empathetic dialogues and adapting to user-specific emotional needs. However, achieving a balance between personalization and safety remains critical, as these tools must operate under ethical constraints to mitigate risks of harmful outputs or dependency [11], [29]. Furthermore, studies evaluating chatbot-assisted mental health therapies highlight promising outcomes in enhancing accessibility for underserved populations, particularly in resource-limited settings where traditional counseling services are scarce [73].

In clinical applications, LLMs have been deployed within healthcare institutions to assist with documentation, diagnostics, and patient engagement. An intriguing case study involved an LLM fine-tuned to summarize clinical notes, reducing the administrative burden on mental health professionals by almost 50

Moreover, LLMs have proven instrumental in community-based mental health initiatives. By deploying multilingual LLMs tailored to regional languages, mental health programs in underserved regions have successfully scaled their interventions, delivering care to populations historically excluded from traditional clinical frameworks [36]. One deployment focused on adolescents exemplifies the capacity of targeted interventions, as LLMs were used to track mood trajectories and deliver tailored therapeutic content based on self-reported emotional states, addressing gaps in adolescent mental health engagement via personalized digital tools [56].

Despite these successes, challenges persist. Real-world applications often encounter issues related to generalizability, ethical and safety considerations, and bias mitigation. For example, domain-specific biases inherent in many pre-trained LLMs result in uneven outcomes for vulnerable populations such as LGBTQ+ youth or underserved ethnic minorities [11]. Furthermore, reliance on noisy or unstructured data, like social media text, introduces risks related to hallucinations or incorrect outputs, necessitating rigorous evaluation and iterative improvements to ensure alignment with clinical standards [38].

In conclusion, these case studies demonstrate the transformative potential of LLMs in mental health services, whether through scalable, automated diagnostics, or accessible counseling tools. However, these deployments also emphasize the necessity of ongoing interdisciplinary collaborations to address bias, safety, and ethical challenges inherent to such systems. Future research should focus on expanding multimodal integration, fostering equity in mental health care, and ensuring sustainable adoption practices that reinforce trust and clinical reliability. By learning from these successful implementations, the field can continue to advance toward more inclusive, impactful, and safe applications.

7.4 Challenges in Model Generalization and Bias

The deployment of large language models (LLMs) in mental health settings encounters critical challenges in generalization, bias, and accuracy, which impede their effectiveness and fairness across diverse populations. Firstly, generalization issues stem from the inherently variable and context-specific nature of mental health interactions. LLMs must function effectively across diverse linguistic, cultural, and socio-demographic contexts, yet many existing models are trained on datasets that are geographically limited or predominantly represent high-resource populations. For instance, datasets like DAIC-WOZ predominantly focus on English-language interactions with specific demographic groups, resulting in limited robustness when applied to conversations from underrepresented cultures or minority languages [43]. Moreover, narrowly sourced training data amplifies biases, particularly in resource-constrained re-

gions where mental health disparities are most pronounced [14].

Bias in LLMs manifests across dimensions such as race, gender, and socio-economic status, leading to unequal outcomes in treatment. Studies highlight that LLMs analyzing mental health conditions often demonstrate reduced empathy or lower diagnostic accuracy for linguistic data from Black users when compared to white users, reflecting imbalances within training datasets [12]. Gender-based disparities are also evident, as masked language models disproportionately associate mental health conditions with women, perpetuating stereotypes while overlooking dimensions of male mental health [77]. Such biases deepen when models encounter multilingual and cross-cultural settings, where mismatches in idioms, sentiment expressions, or syntactic norms limit their reliability in providing accurate or culturally sensitive outputs [78].

In high-stakes contexts, hallucination—when models generate false or misleading outputs—further escalates risks associated with LLM deployment. Erroneous or culturally irrelevant responses to prompts beyond the models' narrowly tuned domains can lead to severe consequences in applications such as early detection of suicidal ideation or automated counseling [75]. While techniques like instruction tuning and domain-specific fine-tuning, as employed by frameworks such as Mental-Alpaca, have effectively mitigated hallucinations, they necessitate extensive domain-relevant datasets, which are often either unavailable or exhibit embedded biases [5].

To address these challenges, enhancing training data diversity and integrating fairness-aware algorithms have emerged as pivotal approaches. Counterfactual data augmentation provides a promising pathway to reduce demographic biases in training corpora [78]. Targeted fine-tuning approaches, exemplified by PsycoLLM's improvements in therapeutic task performance, align models more closely with clinical standards and needs [91]. However, such techniques introduce complexities, including heightened computational requirements and the risk of overfitting on niche subpopulations, thereby reducing broader applicability.

Recent advancements also underscore the importance of explainability and transparency to enhance trust in LLM outputs. Approaches such as dual-prompting and post-hoc explainability frameworks allow stakeholders to better understand how models arrive at their responses, particularly in sensitive scenarios like suicidality detection [112]. Furthermore, cross-cultural adaptation through multilingual training paradigms and fine-tuning with global datasets demonstrates the potential to enhance the applicability of LLMs beyond Western-centric use cases [10].

Addressing the limitations in generalization and biases within LLMs demands a multifaceted strategy that balances data diversification, algorithmic innovation, and rigorous real-world evaluation. Beyond technical solutions, effective interdisciplinary collaboration between AI researchers, clinicians, and ethicists is imperative to ensure equitable and ethical applications of LLMs. As the adoption of AI-driven interventions expands within mental health care, achieving balanced generalization, reducing biases, and improving trustworthiness will be central to enabling more inclusive and impactful solutions globally.

7.5 Iterative Monitoring and Model Improvement

Iterative monitoring and model improvement are indispensable processes for enhancing the performance, safety, and clinical relevance of large language models (LLMs) deployed in mental health applications. This subsection explores frameworks and methodologies for post-deployment evaluation, feedback integration, and domain-specific fine-tuning to ensure that LLMs remain effective and aligned with evolving needs and standards in mental health care.

Effective real-world deployment begins with the establishment of robust post-deployment audit frameworks. These systems are designed to systematically monitor model performance, detect emergent issues, and identify user-related risks through both active and passive feedback channels [83]. For example, usage logs and annotated outputs can be analyzed to detect instances of harmful, biased, or hallucinated responses, which, if unaddressed, could jeopardize user safety. Risk evaluation methods like those detailed in [29] emphasize the importance of continuous auditing to mitigate the instability and potential harms of generative outputs. Tools such as MHaluBench for hallucination detection have similarly been employed in multimodal scenarios to ensure reliability while balancing performance across input modalities [52].

A cornerstone of iterative improvement is the effective collection and utilization of user feedback. Feedback loops have proven to be a critical avenue for dynamic adjustments to model behavior, leveraging insights from users, clinicians, and real-world interactions [66]. Feedback collection can involve post-interaction surveys to assess user satisfaction, surveys gauging therapeutic alignment, or clinician-driven annotations emphasizing the adherence of LLM outputs to clinical mental health standards [66]. Regular comparative analyses of user responses against ground-truth datasets can identify discrepancies in LLM behavior, guiding targeted refinements.

Domain-specific continual pretraining plays an instrumental role in maintaining clinical relevance. Standard LLMs may exhibit initial proficiency in specific mental health tasks; however, they must continually adapt to advances in the field. Approaches such as instruction tuning on curated mental health datasets or incorporating domain-relevant questionnaires, like PHQ-9 or DSM-5 frameworks [49], have demonstrated efficacy in bridging generalized architectures with the nuanced requirements of mental health applications. For instance, Mental-Alpaca and Mental-FLAN-T5, specifically fine-tuned for mental health prediction, showed measurable gains over general-purpose models like GPT-4 [5]. Such approaches underscore the importance of combining generalist capabilities with continuous exposure to clinical data to sustain model reliability and safety.

The alignment of LLMs with advances in mental health research and practice underscores the necessity of adaptive model frameworks. LLMs require regular updates to integrate new insights from clinical research, treatment approaches, and therapeutic methodologies. For instance, multimodal systems incorporating audio or video data alongside text provide a powerful mechanism for tracking complex features indicative of mental health states, such

as prosodic variations and facial microexpressions. Studies such as [17] and [31] highlight how iterative improvement in multimodal capabilities can refine diagnostic precision, thus setting new benchmarks for model performance.

Despite these advancements, significant challenges remain in aligning LLMs with ethical considerations and safety protocols during iterative refinement. Bias detection and mitigation must be integral to model updates, as evidenced by the pervasive impact of demographic imbalances in training datasets [16]. Strategies such as counterfactual data augmentation and fairness-aware fine-tuning have been proposed to address these limitations [16]. Additionally, continuous monitoring systems must adhere to data security and privacy laws, particularly when handling sensitive mental health information [117].

Future research must focus on creating predictive, proactive iterations of LLM frameworks capable of autonomously identifying trends requiring refinement while safeguarding user welfare. Leveraging concepts like temporal event recognition and dynamic adaptation, as explored in [76], can enable LLMs to evolve responsively to emerging user needs without disruptive redeployment cycles.

In essence, iterative monitoring and model improvement ensure that LLMs do not stagnate once deployed but rather evolve continuously to mirror advancements in mental health care. By embedding dynamic feedback loops, leveraging domain-specific datasets, and addressing ethical challenges, LLMs in mental health contexts can maintain a high standard of functionality, safety, and therapeutic relevance over time.

7.6 Ethical, Privacy, and Safety Considerations in Evaluation

The evaluation and deployment of Large Language Models (LLMs) for mental health services introduce profound ethical, privacy, and safety considerations, particularly given the sensitive nature of mental health data and the high-stakes implications of model outputs. Following the previous discussion on iterative improvement, this subsection delves into pivotal challenges and emerging strategies to safeguard privacy, ensure ethical adherence, and mitigate risks during model assessment and real-world use.

A foundational ethical concern in LLM-based mental health applications is the preservation of privacy and confidentiality of patient data. Mental health records contain inherently sensitive information, and the potential for reidentification—via nuanced linguistic patterns or contextual elements in model outputs—amplifies privacy risks. Traditional anonymization techniques often fail to comprehensively address these challenges, as mental health information can be deeply personal and implicitly revealing. Techniques like differential privacy, which apply controlled noise to protect individual contributions, have shown promise but may trade off some degree of model performance [39]. To mitigate these risks further, privacy-first strategies utilizing open-source LLMs fine-tuned in controlled environments have been proposed to reduce reliance on large-scale corporate pipelines that centralize sensitive data [5].

Beyond privacy, the safety of LLM responses in mental health contexts is an equally critical focus to prevent harmful or inappropriate outputs during user interactions. Misdiagnoses, harmful therapeutic suggestions, or triggering content can endanger vulnerable users relying on these models. While advanced LLMs, such as GPT-4, have demonstrated their ability to produce empathetic interactions, issues like hallucination—where outputs diverge significantly from factual or clinically validated information—persist as a key concern [100]. To address this, robust pre-deployment evaluation protocols, including adversarial stress testing and safety-targeted assessments, have been advocated, especially for high-risk scenarios like detecting suicidal ideation [68]. Moreover, integrating real-time monitoring systems with automated alerts for human intervention can enhance user safety in live interactions [68].

Transparency in LLM decision-making further bolsters ethical applicability, as the current opacity of these systems hinders interpretability and trust. Explainable AI (XAI) techniques, such as saliency maps and attribution models, offer promising tools to contextualize outputs and identify errors, thus enabling validation against evidence-based clinical standards [97]. Clearly presented rationales for therapeutic guidance can bridge gaps between technological outputs and practical clinical use, fostering trust among clinicians and end-users alike [66]. Structured ethical guidelines, co-created by interdisciplinary teams of clinicians, technologists, and ethicists, are essential to ensure consistency with evidence-based mental health practices [12].

Compliance with legal and regulatory standards introduces another layer of complexity. LLM-based systems must align with established frameworks like HIPAA, GDPR, and equivalent regulations, ensuring lawful data handling and minimizing liabilities. Meticulous documentation and comprehensive auditing practices have been suggested to reinforce accountability, particularly for real-world applications [69]. Regulatory compliance must also be complemented by human oversight mechanisms and iterative validation processes to align LLM outputs with both societal safety concerns and clinical norms [118].

Despite advancements in these domains, emerging ethical challenges compel further exploration. Biases embedded in LLM training data can lead to significant disparities in how mental health services are delivered to diverse populations. For instance, demographic or cultural biases might influence how empathetically an LLM responds to individuals from different backgrounds, highlighting the necessity for fairness-aware model development and evaluation [12]. Furthermore, over-reliance on LLMs poses risks of delaying critical professional intervention in high-stakes cases, raising ethical dilemmas regarding the balance between automation and human expertise [67].

In conclusion, while LLMs represent a transformative opportunity for mental health services, addressing ethical, privacy, and safety considerations is paramount for their responsible evaluation and deployment. By leveraging privacy-preserving mechanisms, implementing harm-reduction strategies, adopting explainable AI tools, and adhering to interdisciplinary regulatory frameworks, these challenges can be mitigated. Building on iterative refinement strategies and looking toward future advancements, maintaining empirical validation and ethical rigor will ensure that LLMs support, rather than undermine, the com-

plex and sensitive ecosystem of mental health care [69].

8 Future Directions and Conclusion

The transformative potential of Large Language Models (LLMs) in mental health services has been unveiled through their expansive capabilities in diagnostics, therapeutic interventions, accessibility enhancement, and professional support. However, realizing their full utility requires addressing complex challenges, advancing interdisciplinary methodologies, and upholding rigorous ethical principles to ensure their responsible deployment. This section synthesizes the state of the field and identifies future directions to propel the integration of LLMs into mental health systems.

The integration of multimodal data represents a critical frontier in enhancing the precision and scope of mental health assessments. Exploiting rich streams of information from diverse modalities—such as textual data, speech, physiological signals (e.g., EEG), and facial expressions—promises to deepen insights into psychological states. Emerging models that combine these inputs demonstrate gains in diagnostic accuracy and emotional understanding [13], [17]. However, challenges persist in harmonizing disparate data sources and overcoming computational burdens, emphasizing the need for innovative architectures optimized for multimodal integration. Future research could explore synergizing domain-specific pretraining with low-resource-efficient techniques, ensuring these methods remain scalable across varied settings.

Cross-cultural and linguistic adaptability is an essential direction to ensure that LLMs meet the mental health needs of diverse populations worldwide. Despite advancements, current models often exhibit biases or limitations when addressing non-dominant languages and culturally grounded expressions of mental health [16], [86]. The development of culturally aware LLMs, trained on underrepresented datasets and guided by interdisciplinary collaborations with sociolinguists and mental health professionals, is critical to reducing global disparities in mental health care. Techniques such as counterfactual data augmentation and fairness-aware model fine-tuning can bolster system inclusivity [11], [95].

Explainability has emerged as a cornerstone for building trust in LLM deployment within sensitive applications like mental health. Ensuring that outputs align with clinical principles while remaining transparent to users and professionals is critical to their acceptance in practice [63]. Promising advancements in explainable AI methodologies, including rationale extraction and attribution systems, underscore the potential for models to present both outcomes and the decision-making processes behind their recommendations. Nonetheless, balancing these technical implementations with usability remains complex, calling for more granular evaluations incorporating both subjective (e.g., user trust) and objective (e.g., diagnostic alignment) metrics [63], [119].

The societal and systemic implications of broad LLM adoption in mental health merit deeper scrutiny, particularly regarding sustained user engagement, potential over-reliance, and evolving practitioner-patient dynamics. As demonstrated by pilot studies leveraging conversational agents [74], [101], while users benefit from instant, judgment-free interactions, the risk of dependency or misaligned expectations remains pertinent. Future investigations need to address how LLM-powered tools complement human counselors rather than act as substitutes, preserving therapeutic boundaries while enhancing workflows. Such efforts should incorporate empirical longitudinal studies to monitor the lasting impact of LLM-based mental health tools on users' well-being and clinical outcomes.

Finally, advancing responsible LLM innovation demands robust interdisciplinary collaboration. Policymakers, data scientists, ethicists, and mental health experts must unite to design clear frameworks governing the safe and equitable use of LLMs [11], [69]. Data privacy, fairness, and accountability remain paramount to prevent misuse and build systems rooted in ethical integrity. Emphasizing the integration of user-informed consent mechanisms and regulatory oversight can further underpin trust and societal acceptance.

In conclusion, LLMs hold transformative potential for reshaping mental health services, but their adoption requires surmounting technical, cultural, and ethical barriers. Their future lies in multimodal enhancement, contextual sensitivity, explainable outputs, and responsible deployment. By fostering sustained collaboration across academia, industry, and public health sectors, LLMs can be judiciously leveraged to address unmet mental health needs while adhering to principles of equity, inclusivity, and safety.

References

[1] L. Liu, X. Yang, J. Lei, X. Liu, Y. Shen, Z. Zhang, P. Wei, J. Gu, Z. Chu, Z. Qin, and K. Ren, "A survey on medical large language models: Technology, application, trustworthiness, and future directions," *ArXiv*, vol. abs/2406.03712, 2024. 1, 14, 19, 20

[2] S. Ji, T. Zhang, L. Ansari, J. Fu, P. Tiwari, and E. Cambria, "Mentalbert: Publicly available pretrained language models for mental healthcare," *ArXiv*, vol. abs/2110.15621, 2021. 1, 2, 10, 23

[3] K. He, R. Mao, Q. Lin, Y. Ruan, X. Lan, M. Feng, and E. Cambria, "A survey of large language models for healthcare: from data, technology, and applications to accountability and ethics," *ArXiv*, vol. abs/2310.05694, 2023. 1, 15

[4] A. G. Reece, A. J. Reagan, K. Lix, P. Dodds, C. Danforth, and E. Langer, "Forecasting the onset and course of mental illness with twitter data," *Scientific Reports*, vol. 7, 2016. 1, 2, 3, 13, 23

[5] X. Xu, B. Yao, Y. Dong, H. Yu, J. A. Hendler, A. Dey, and D. Wang, "Leveraging large language models for mental health prediction via online text data," *ArXiv*, vol. abs/2307.14385, 2023. 1, 2, 3, 6, 7, 8, 12, 14, 15, 16, 20, 22, 25, 26, 27

[6] E. P. Lehman, E. Hernandez, D. Mahajan, J. Wulff, M. J. Smith, Z. M. Ziegler, D. Nadler, P. Szolovits, A. E. W. Johnson, and E. Alsentzer, "Do we still need clinical language models?" *ArXiv*, vol. abs/2302.08091, 2023. 1

[7] C. Wang, M. Li, J. He, Z. Wang, E. Darzi, Z. Chen, J. Ye, T.-X. Li, Y.-C. Su, J. Ke, K. Qu, S. Li, Y. Yu, P. Liò, T. Wang, Y. G. Wang, and Y. Shen, "A survey for large language models in biomedicine," *ArXiv*, vol. abs/2409.00133, 2024. 1, 2, 18, 19

[8] H. Na, "Cbt-llm: A chinese large language model for cognitive behavioral therapy-based mental health question answering," in *International Conference on Language Resources and Evaluation*, 2024, pp. 2930–2940. 1, 6, 10

[9] Y. Y. Chiu, A. Sharma, I. W. Lin, and T. Althoff, "A computational framework for behavioral assessment of llm therapists," *ArXiv*, vol. abs/2401.00820, 2024. 1, 9, 12, 14, 15, 23

[10] T. Lai, Y. Shi, Z. Du, J. Wu, K. Fu, Y. Dou, and Z. Wang, "Psy-llm: Scaling up global mental health psychological services with ai-based large language models," *ArXiv*, vol. abs/2307.11991, 2023. 1, 5, 6, 8, 9, 10, 12, 14, 15, 18, 23, 26

[11] H. R. Lawrence, R. A. Schneider, S. B. Rubin, M. J. Mataric, D. McDuff, and M. J. Bell, "The opportunities and risks of large language models in mental health," *JMIR Mental Health*, vol. 11, 2024. 2, 3, 6, 13, 14, 16, 17, 24, 25, 28

[12] S. Gabriel, I. Puri, X. Xu, M. Malgaroli, and M. Ghassemi, "Can ai relate: Testing large language model response for mental health support," *ArXiv*, vol. abs/2405.12021, 2024. 2, 11, 14, 15, 16, 20, 23, 24, 26, 27

[13] Y. Hu, S. Zhang, T. Dang, H. Jia, F. D. Salim, W. Hu, and A. Quigley, "Exploring large-scale language models to evaluate eeg-based multimodal data for mental health," *ArXiv*, vol. abs/2408.07313, 2024. 2, 3, 5, 6, 10, 11, 13, 15, 16, 23, 24, 28

[14] A. Cohan, B. Desmet, A. Yates, L. Soldaini, S. MacAvaney, and N. Goharian, "Smhd: a large-scale resource for exploring online language usage for multiple mental health conditions," *ArXiv*, vol. abs/1806.05258, 2018. 2, 3, 5, 8, 12, 13, 15, 16, 17, 23, 26

[15] B. Lamichhane, "Evaluation of chatgpt for nlp-based mental health applications," *ArXiv*, vol. abs/2303.15727, 2023. 2, 8, 9, 15, 16, 18, 20, 24

[16] A. Murarka, B. Radhakrishnan, and S. Ravichandran, "Detection and classification of mental illnesses on social media using roberta," *ArXiv*, vol. abs/2011.11226, 2020. 2, 5, 13, 15, 19, 27, 28

[17] A. Haque, M. Guo, A. S. Miner, and L. Fei-Fei, "Measuring depression symptom severity from spoken language and 3d facial expressions," *ArXiv*, vol. abs/1811.08592, 2018. 2, 4, 5, 8, 9, 11, 13, 18, 20, 21, 23, 27, 28

[18] L. Zhang, X. Huang, T. Liu, Z. Chen, and T. Zhu, "Using linguistic features to estimate suicide probability of chinese microblog users," *ArXiv*, vol. abs/1411.0861, 2014. 3, 20

[19] X. Chen, M. Sykora, T. W. Jackson, and S. Elayan, "What about mood swings: Identifying depression on twitter with temporal measures of emotions," *Companion Proceedings of the The Web Conference 2018*, 2018. 3, 11, 19

[20] M. Trotzek, S. Koitka, and C. Friedrich, "Utilizing neural networks and linguistic metadata for early detection of depression indications in text sequences," *IEEE Transactions on Knowledge and Data Engineering*, vol. 32, pp. 588–601, 2018. 3

[21] I. Sekulic and M. Strube, "Adapting deep learning methods for mental health prediction on social media," *ArXiv*, vol. abs/2003.07634, 2019. 3, 15

[22] S. Ji, X. Li, Z. Huang, and E. Cambria, "Suicidal ideation and mental disorder detection with attentive relation networks," *Neural Computing and Applications*, vol. 34, pp. 10 309 – 10 319, 2020. 3, 4, 19

[23] A. Kumar, A. Sharma, and A. Arora, "Anxious depression prediction in real-time social data," *MatSciRN: Other Biomaterials (Topic)*, 2019. 3

[24] S. Yadav, J. Chauhan, J. P. Sain, K. Thirunarayan, A. Sheth, and J. A. Schumm, "Identifying depressive symptoms from tweets: Figurative language enabled multitask learning framework," in *International Conference on Computational Linguistics*, 2020, pp. 696–709. 3, 5, 13, 20

[25] M. Kabir, T. Ahmed, M. B. Hasan, M. T. R. Laskar, T. K. Joarder, H. Mahmud, and K. Hasan, "Deptweet: A typology for social media texts to detect depression severities," *Comput. Hum. Behav.*, vol. 139, p. 107503, 2022. 3

[26] G. Rosenman, L. Wolf, and T. Hendler, "Llm questionnaire completion for automatic psychiatric assessment," *ArXiv*, vol. abs/2406.06636, 2024. 3, 11

[27] W. Qin, Z. Chen, L. Wang, Y. Lan, W. Ren, and R. Hong, "Read, diagnose and chat: Towards explainable and interactive llms-augmented depression detection in social media," *ArXiv*, vol. abs/2305.05138, 2023. 3, 9, 13, 18, 21, 24

[28] N. C. Chung, G. C. Dyer, and L. Brocki, "Challenges of large language models for mental health counseling," *ArXiv*, vol. abs/2311.13857, 2023. 3, 7, 11, 19, 20

[29] M. D. Choudhury, S. R. Pendse, and N. Kumar, "Benefits and harms of large language models in digital mental health," *ArXiv*, vol. abs/2311.14693, 2023. 3, 4, 7, 8, 9, 12, 18, 20, 21, 22, 25, 26

[30] S. MacAvaney, B. Desmet, A. Cohan, L. Soldaini, A. Yates, A. Zirikly, and N. Goharian, "Rsdd-time: Temporal annotation of self-reported mental health diagnoses," *ArXiv*, vol. abs/1806.07916, 2018. 3, 4, 5, 9, 12, 16, 20

[31] X. Zhang, H. Liu, K. Xu, Q. Zhang, D. Liu, B. Ahmed, and J. Epps, "When llms meets acoustic landmarks: An efficient approach to integrate speech into large language models for depression detection," *ArXiv*, vol. abs/2402.13276, 2024. 3, 5, 7, 9, 13, 18, 20, 27

[32] I. Galatzer-Levy, D. J. McDuff, V. Natarajan, A. Karthikesalingam, and M. Malgaroli, "The capability of large language models to measure psychiatric functioning," *ArXiv*, vol. abs/2308.01834, 2023. 3, 12

[33] A. Laverghetta, A. Nighojkar, J. Mirzakhalov, and J. Licato, "Predicting human psychometric properties using computational language models," *ArXiv*, vol. abs/2205.06203, 2022. 3

[34] A. Mitra, R. Pradhan, R. Melamed, K. Chen, D. Hoaglin, K. Tucker, J. Reisman, Z. Yang, W. Liu, J. Tsai, and H. Yu, "Associations between natural language processing–enriched social determinants of health and suicide death among us veterans," *JAMA Network Open*, vol. 6, 2022. 3, 8

[35] X. Lan, Y. Cheng, L. Sheng, C. Gao, and Y. Li, "Depression detection on social media with large language models," *ArXiv*, vol. abs/2403.10750, 2024. 3, 11, 16, 20, 25

[36] E. A. R'issola, M. Aliannejadi, and F. Crestani, "Mental disorders on online social media through the lens of language and behaviour: Analysis and visualisation," *ArXiv*, vol. abs/2202.03291, 2022. 3, 8, 11, 16, 20, 25

[37] A. Yates, A. Cohan, and N. Goharian, "Depression and self-harm risk assessment in online forums," in *Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 2968–2978. 4

[38] K. Yang, T. Zhang, Z.-Z. Kuang, Q. Xie, and S. Ananiadou, "Mentallama: Interpretable mental health analysis on social media with large language models," *Proceedings of the ACM on Web Conference 2024*, 2023. 4, 11, 12, 21, 25

[39] L. Li, J. Zhou, Z. Gao, W. Hua, L. Fan, H. Yu, L. Hagen, Y. Zhang, T. L. Assimes, L. Hemphill, and S. Ma, "A scoping review of using large language models (llms) to investigate electronic health records (ehrs)," *ArXiv*, vol. abs/2405.03066, 2024. 4, 27

[40] A. Belyaeva, J. Cosentino, F. Hormozdiari, K. Eswaran, S. Shetty, G. C. Corrado, A. Carroll, C. Y. McLean, and N. Furlotte, "Multimodal llms for health grounded in individual-specific data," in *ML4MHD*, 2023, pp. 86–102. 4, 8, 12, 17

[41] K. Roy, M. Gaur, M. Soltani, V. Rawte, A. Kalyan, and A. P. Sheth, "Proknow: Process knowledge for safety constrained and explainable question generation for mental health diagnostic assistance," *Frontiers in Big Data*, vol. 5, 2023. 4, 12

[42] P. K. Kanithi, C. Christophe, M. A. Pimentel, T. Raha, N. Saadi, H. Javed, S. Maslenkova, N. Hayat, R. Rajan, and S. Khan, "Medic: Towards a comprehensive framework for evaluating llms in clinical applications," *ArXiv*, vol. abs/2409.07314, 2024. 4, 17, 21

[43] S. Burdisso, E. Reyes-Ram'irez, E. Villatoro-Tello, F. S'anchez-Vega, A. P. Lopez-Monroy, and P. Motlícek, "Daic-woz: On the validity of using the therapist's prompts in automatic depression detection from clinical interviews," *ArXiv*, vol. abs/2404.14463, 2024. 4, 5, 25

[44] Z. Chen, P. Kulkarni, I. Galatzer-Levy, B. Bigio, C. Nasca, and Y. Zhang, "Modern views of machine learning for precision psychiatry," *Patterns*, vol. 3, 2022. 4

[45] C. Flint, M. Cearns, N. Opel, R. Redlich, D. Mehler, D. Emden, N. Winter, R. Leenings, S. Eickhoff, T. Kircher, A. Krug, I. Nenadić, V. Arolt, S. Clark, B. Baune, X. Jiang, U. Dannlowski, and T. Hahn, "Systematic misestimation of machine learning performance in neuroimaging studies of depression," *Neuropsychopharmacology*, vol. 46, pp. 1510 – 1517, 2019. 5

[46] B. Yao, C. Shi, L. Zou, L. Dai, M. Wu, L. Chen, Z. Wang, and K.-M. Yu, "D4: a chinese dialogue dataset for depression-diagnosis-oriented chat," *ArXiv*, vol. abs/2205.11764, 2022. 5, 21

[47] L. Weidinger, J. F. J. Mellor, M. Rauh, C. Griffin, J. Uesato, P.-S. Huang, M. Cheng, M. Glaese, B. Balle, A. Kasirzadeh, Z. Kenton, S. Brown, W. Hawkins, T. Stepleton, C. Biles, A. Birhane, J. Haas, L. Rimell, L. A. Hendricks, W. S. Isaac, S. Legassick, G. Irving, and I. Gabriel, "Ethical and social risks of harm from language models," *ArXiv*, vol. abs/2112.04359, 2021. 5

[48] K. Mei, S. Fereidooni, and A. Caliskan, "Bias against 93 stigmatized groups in masked language models and downstream sentiment classification tasks," *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 2023. 5, 17

[49] T. Nguyen, A. Yates, A. Zirikly, B. Desmet, and A. Cohan, "Improving the generalizability of depression detection by leveraging clinical questionnaires," in *Annual Meeting of the Association for Computational Linguistics*, 2022, pp. 8446–8459. 5, 9, 13, 14, 26

[50] P. Wei, K. Peng, A. Roitberg, K. Yang, J. Zhang, and R. Stiefelhagen, "Multi-modal depression estimation based on sub-attentional fusion," in *ECCV Workshops*, 2022, pp. 623–639. 5

[51] Z. Chen, R. Yang, S. Fu, N. Zong, H. Liu, and M. Huang, "Detecting reddit users with depression using a hybrid neural network sbert-cnn," *2023 IEEE 11th International Conference on Healthcare Informatics (ICHI)*, pp. 193–199, 2023. 5, 9

[52] X. Chen, C. Wang, Y. Xue, N. Zhang, X. Yang, Q. Li, Y. Shen, L. Liang, J. Gu, and H. Chen, "Unified hallucination detection for multimodal large language models," *ArXiv*, vol. abs/2402.03190, 2024. 5, 18, 26

[53] L. Ilias, S. Mouzakitis, and D. Askounis, "Calibration of transformer-based models for identifying stress and depression in social media," *IEEE Transactions on Computational Social Systems*, vol. 11, pp. 1979–1990, 2023. 5, 15

[54] W. Santos and I. Paraboni, "Prompt-based mental health screening from social media text," *ArXiv*, vol. abs/2401.05912, 2024. 5

[55] A.-M. Bucur, A. Cosma, P. Rosso, and L. P. Dinu, "It's just a matter of time: Detecting depression with time-enriched multimodal transformers," in *European Conference on Information Retrieval*, 2023, pp. 200–215. 5

[56] I. Ameer, M. Arif, G. Sidorov, H. Gómez-Adorno, and A. Gelbukh, "Mental illness classification on social media texts using deep learning and transfer learning," *ArXiv*, vol. abs/2207.01012, 2022. 5, 25

[57] G. Wang, G. Yang, Z. Du, L. Fan, and X. Li, "Clinicalgpt: Large language models finetuned with diverse medical data and comprehensive evaluation," *ArXiv*, vol. abs/2306.09968, 2023. 6, 10, 23

[58] C. Christophe, P. K. Kanithi, T. Raha, S. Khan, and M. A. Pimentel, "Med42-v2: A suite of clinical llms," *ArXiv*, vol. abs/2408.06142, 2024. 6

[59] A. Sharma, A. S. Miner, D. C. Atkins, and T. Althoff, "A computational approach to understanding empathy expressed in text-based mental health support," *ArXiv*, vol. abs/2009.08441, 2020. 6, 23

[60] H. Zhan, A. Zheng, Y. K. Lee, J. Suh, J. J. Li, and D. C. Ong, "Large language models are capable of offering cognitive reappraisal, if guided," *ArXiv*, vol. abs/2404.01288, 2024. 6, 15, 23

[61] D. Dash, R. Thapa, J. Banda, A. Swaminathan, M. Cheatham, M. Kashyap, N. Kotecha, J. H. Chen, S. Gombar, L. Downing, R. A. Pedreira, E. Goh, A. Arnaout, G. K. Morris, H. Magon, M. Lungren, E. Horvitz, and N. Shah, "Evaluation of gpt-3.5 and gpt-4 for supporting real-world information needs in healthcare delivery," *ArXiv*, vol. abs/2304.13714, 2023. 6, 10

[62] Z. Bao, W. Chen, S. Xiao, K. Ren, J. Wu, C. Zhong, J. Peng, X. Huang, and Z. Wei, "Disc-medllm: Bridging general large language models and real-world medical consultation," *ArXiv*, vol. abs/2308.14346, 2023. 6, 10

[63] K. Yang, S. Ji, T. Zhang, Q. Xie, Z.-Z. Kuang, and S. Ananiadou, "Towards interpretable mental health analysis with large language models," in *Conference on Empirical Methods in Natural Language Processing*, 2023, pp. 6056–6077. 6, 19, 28

[64] A. Sharma, K. Rushton, I. W. Lin, D. Wadden, K. G. Lucas, A. S. Miner, T. Nguyen, and T. Althoff, "Cognitive reframing of negative thoughts through human-language model interaction," in *Annual Meeting of the Association for Computational Linguistics*, 2023, pp. 9977–10 000. 7, 20

[65] A. Sharma, K. Rushton, I. W. Lin, T. Nguyen, and T. Althoff, "Facilitating self-guided mental health interventions through human-language model interaction: A case study of cognitive restructuring," *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2023. 7, 15

[66] P. K. Adhikary, A. Srivastava, S. Kumar, S. M. Singh, P. Manuja, J. K. Gopinath, V. Krishnan, S. Kedia, K. Deb, and T. Chakraborty, "Exploring the efficacy of large language models in summarizing mental health counseling sessions: Benchmark study," *JMIR Mental Health*, vol. 11, 2024. 7, 13, 18, 26, 27

[67] Z. Ma, Y. Mei, and Z. Su, "Understanding the benefits and challenges of using large language model-based conversational agents for mental well-being support," *AMIA ... Annual Symposium proceedings. AMIA Symposium*, vol. 2023, pp. 1105–1114, 2023. 7, 8, 17, 19, 24, 27

[68] J. I. Park, M. Abbasian, I. Azimi, D. Bounds, A. Jun, J. Han, R. McCarron, J. Borelli, J. Li, M. Mahmoudi, C. Wiedenhoeft, and A. Rahmani, "Building trust in mental health chatbots: Safety metrics and llm-based evaluation tools," *ArXiv*, vol. abs/2408.04650, 2024. 7, 8, 9, 13, 17, 20, 21, 22, 24, 27

[69] Y. Hua, F. Liu, K. Yang, Z. Li, H. Na, Y. Sheu, P. Zhou, L. V. Moran, S. Ananiadou, and A. Beam, "Large language models in mental health care: a scoping review," *ArXiv*, vol. abs/2401.02984, 2024. 7, 11, 13, 14, 19, 20, 21, 22, 27, 28

[70] X. Chen, J. Xiang, S. Lu, Y. Liu, M. He, and D. Shi, "Evaluating large language models in medical applications: a survey," *ArXiv*, vol. abs/2405.07468, 2024. 7

[71] Y. Kim, X. Xu, D. McDuff, C. Breazeal, and H. W. Park, "Health-llm: Large language models for health prediction via wearable sensor data," *ArXiv*, vol. abs/2401.06866, 2024. 7, 16

[72] Y. Huang, K. Tang, and M. Chen, "A comprehensive survey on evaluating large language model applications in the medical industry," *ArXiv*, vol. abs/2404.15777, 2024. 7, 8, 20, 23

[73] S. Lissak, N. Calderon, G. Shenkman, Y. Ophir, E. Fruchter, A. Klomek, and R. Reichart, "The colorful future of llms: Evaluating and improving llms as emotional supporters for queer youth," in *North American Chapter of the Association for Computational Linguistics*, 2024, pp. 2040–2079. 8, 20, 25

[74] I. Song, S. R. Pendse, N. Kumar, and M. D. Choudhury, "The typing cure: Experiences with large language model chatbots for mental health support," *ArXiv*, vol. abs/2401.14362, 2024. 8, 17, 21, 24, 28

[75] M. A. Ahmad, I. Yaramis, and T. D. Roy, "Creating trustworthy llms: Dealing with hallucinations in healthcare ai," *ArXiv*, vol. abs/2311.01463, 2023. 8, 12, 17, 21, 26

[76] M. Gaur, V. Aribandi, A. Alambo, U. Kursuncu, K. Thirunarayan, J. Beich, J. Pathak, and A. Sheth, "Characterization of time-variant and time-invariant assessment of suicidality on reddit using c-ssrs," *PLoS ONE*, vol. 16, 2021. 8, 21, 25, 27

[77] I. W. Lin, L. Njoo, A. Field, A. Sharma, K. Reinecke, T. Althoff, and Y. Tsvetkov, "Gendered mental health stigma in masked language models," in *Conference on Empirical Methods in Natural Language Processing*, 2022, pp. 2152–2170. 9, 17, 26

[78] Y. Wang, Y. Zhao, S. A. Keller, A. D. Hond, M. M. van Buchem, M. Pillai, and T. Hernandez-Boussard, "Unveiling and mitigating bias in mental health analysis with large language models," *ArXiv*, vol. abs/2406.12033, 2024. 9, 17, 26

[79] N. Flemotomos, V. R. Martinez, Z. Chen, K. Singla, V. Ardulov, R. Peri, D. D. Caperton, J. Gibson, M. J. Tanana, P. Georgiou, J. V. Epps, S. P. Lord, T. Hirsch, Z. E. Imel, D. C. Atkins, and S. S. Narayanan, "Automated evaluation of psychotherapy skills using speech and language technologies," *Behavior Research Methods*, vol. 54, pp. 690 – 711, 2021. 9

[80] A. Ray, S. Kumar, R. Reddy, P. Mukherjee, and R. Garg, "Multi-level attention network using text, audio and video for depression prediction," *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, 2019. 9

[81] B. Lin, D. Bouneffouf, G. Cecchi, and R. Tejwani, "Neural topic modeling of psychotherapy sessions," *ArXiv*, vol. abs/2204.10189, 2022. 9

[82] T. Althoff, K. Clark, and J. Leskovec, "Large-scale analysis of counseling conversations: An application of natural language processing to mental health," *Transactions of the Association for Computational Linguistics*, vol. 4, pp. 463 – 476, 2016. 9

[83] S. Ji, T. Zhang, K. Yang, S. Ananiadou, and E. Cambria, "Rethinking large language models in mental health applications," *ArXiv*, vol. abs/2311.11267, 2023. 9, 11, 13, 18, 22, 24, 26

[84] D. V. Veen, C. V. Uden, L. Blankemeier, J.-B. Delbrouck, A. Aali, C. Blüthgen, A. Pareek, M. Polacin, W. Collins, N. Ahuja, C. Langlotz, J. Hom, S. Gatidis, J. M. Pauly, and A. S. Chaudhari, "Adapted large language models can outperform medical experts in clinical text summarization." *Nature medicine*, 2023. 10

[85] X. Yang, N. M. Pournejatian, H.-C. Shin, K. E. Smith, C. Parisien, C. B. Compas, C. Martin, M. G. Flores, Y. Zhang, T. Magoc, C. Harle, G. P. Lipori, D. A. Mitchell, W. Hogan, E. Shenkman, J. Bian, and Y. Wu, "Gatortron: A large clinical language model to unlock patient information from unstructured electronic health records," *ArXiv*, vol. abs/2203.03540, 2022. 10, 19, 23

[86] W. Zhai, H. Qi, Q. Zhao, J. Li, Z. Wang, H. Wang, B. X. Yang, and G. Fu, "Chinese mentalbert: Domain-adaptive pre-training on social media for chinese mental health text analysis," in *Annual Meeting of the Association for Computational Linguistics*, 2024, pp. 10 574–10 585. 10, 24, 28

[87] I. Jahan, M. T. R. Laskar, C. Peng, and J. X. Huang, "A comprehensive evaluation of large language models on benchmark biomed-

ical text processing tasks," *Computers in biology and medicine*, vol. 171, p. 108189, 2023. 10

[88] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *International Conference on Learning Representations*, 2013. 11

[89] W. Street, J. O. Siy, G. Keeling, A. Baranes, B. Barnett, M. McKibben, T. Kanyere, A. Lentz, B. A. Y. Arcas, and R. I. M. Dunbar, "Llms achieve adult human performance on higher-order theory of mind tasks," *ArXiv*, vol. abs/2405.18870, 2024. 11

[90] T. Kim, S. Bae, H. A. Kim, S.-W. Lee, H. Hong, C. Yang, and Y.-H. Kim, "Mindfuldiary: Harnessing large language model to support psychiatric patients' journaling," *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2023. 12

[91] J. Hu, T. Dong, H. Ma, P. Zou, X. Sun, and M. Wang, "Psycollm: Enhancing llm for psychological understanding and evaluation," *ArXiv*, vol. abs/2407.05721, 2024. 12, 21, 26

[92] P. M. Doraiswamy, C. Blease, and K. A. Bodner, "Artificial intelligence and the future of psychiatry: Insights from a global physician survey," *Artificial intelligence in medicine*, vol. 102, p. 101753, 2019. 13

[93] E. Berberette, J. Hutchins, and A. Sadovnik, "Redefining "hallucination" in llms: Towards a psychology-informed framework for mitigating misinformation," *ArXiv*, vol. abs/2402.01769, 2024. 13, 21

[94] T. Zhang, S. Teng, H. Jia, and S. D'Alfonso, "Leveraging llms to predict affective states via smartphone sensor features," *ArXiv*, vol. abs/2407.08240, 2024. 13, 17

[95] Z. Guo, A. Lai, J. H. Thygesen, J. Farrington, T. Keen, and K. Li, "Large language model for mental health: A systematic review," *ArXiv*, vol. abs/2403.15401, 2024. 13, 28

[96] S. C. Guntuku, A. Buffone, K. Jaidka, J. Eichstaedt, and L. Ungar, "Understanding and measuring psychological stress using social media," *ArXiv*, vol. abs/1811.07430, 2018. 13, 15

[97] M. T. R. Laskar, S. Alqahtani, M. S. Bari, M. Rahman, M. A. M. Khan, H. Khan, I. Jahan, A. Bhuiyan, C. W. Tan, M. R. Parvez, E. Hoque, S. R. Joty, and J. X. Huang, "A systematic survey and critical review on evaluating large language models: Challenges, limitations, and recommendations," *ArXiv*, vol. abs/2407.04069, 2024. 14, 23, 27

[98] B. Abeysinghe and R. Circi, "The challenges of evaluating llm applications: An analysis of automated, human, and llm-based approaches," *ArXiv*, vol. abs/2406.03339, 2024. 14

[99] G. Fu, Q. Zhao, J. Li, D. Luo, C. Song, W. Zhai, S. Liu, F. Wang, Y. Wang, L. Cheng, J. Zhang, and B. Yang, "Enhancing psychological counseling with large language model: A multifaceted decision-support system for non-professionals," *ArXiv*, vol. abs/2308.15192, 2023. 14

[100] M. Luo, C. J. Warren, L. Cheng, H. M. Abdul-Muhsin, and I. Banerjee, "Assessing empathy in large language models with real-world physician-patient interactions," *ArXiv*, vol. abs/2405.16402, 2024. 14, 22, 27

[101] A. Singh, A. Ehtesham, S. Mahmud, and J.-H. Kim, "Revolutionizing mental health care through langchain: A journey with a large language model," *2024 IEEE 14th Annual Computing and Communication Workshop and Conference (CCWC)*, pp. 0073–0078, 2024. 15, 28

[102] H. Zhou, B. Gu, X. Zou, Y. Li, S. S. Chen, P. Zhou, J. Liu, Y. Hua, C. Mao, X. Wu, Z. Li, and F. Liu, "A survey of large language models in medicine: Progress, application, and challenge," *ArXiv*, vol. abs/2311.05112, 2023. 15

[103] A. Tamkin, A. Askell, L. Lovitt, E. Durmus, N. Joseph, S. Kravec, K. Nguyen, J. Kaplan, and D. Ganguli, "Evaluating and mitigating discrimination in language model decisions," *ArXiv*, vol. abs/2312.03689, 2023. 17, 21

[104] D. Tao, H. Chui, S. Luk, and T. Lee, "Cuempathy: A counseling speech dataset for psychotherapy research," *2022 13th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pp. 354–358, 2022. 18

[105] Z. Shi, Z. Wang, H. Fan, Z. Zhang, L. Li, Y. Zhang, Z. fei Yin, L. Sheng, Y. Qiao, and J. Shao, "Assessment of multimodal large language models in alignment with human values," *ArXiv*, vol. abs/2403.17830, 2024. 18

[106] G. Han, W. Liu, X. Huang, and B. Borsari, "Chain-of-interaction: Enhancing large language models for psychiatric behavior understanding by dyadic contexts," *2024 IEEE 12th International Conference on Healthcare Informatics (ICHI)*, pp. 392–401, 2024. 18, 22

[107] K. Harrigian, C. A. Aguirre, and M. Dredze, "On the state of social media data for mental health research," *ArXiv*, vol. abs/2011.05233, 2020. 18

[108] E. C. Acikgoz, O. B. .Ince, R. Bench, A. A. Boz, I. Kesen, A. Erdem, and E. Erdem, "Hippocrates: An open-source framework for advancing large language models in healthcare," *ArXiv*, vol. abs/2404.16621, 2024. 19

[109] S. Wo'zniak, B. Koptyra, A. Janz, P. Kazienko, and J. Koco'n, "Personalized large language models," *ArXiv*, vol. abs/2402.09269, 2024. 19

[110] S. Schmidgall, C. Harris, I. Essien, D. Olshvang, T. Rahman, J. W. Kim, R. Ziaei, J. Eshraghian, P. M. Abadir, and R. Chellappa, "Addressing cognitive bias in medical language models," *ArXiv*, vol. abs/2402.08113, 2024. 20

[111] Z. Zhang, S. Chen, M. Wu, and K. Zhu, "Psychiatric scale guided risky post screening for early detection of depression," in *International Joint Conference on Artificial Intelligence*, 2022, pp. 5220–5226. 21, 25

[112] H. Jeon, D. Yoo, D. Lee, S. Son, S. Kim, and J. Han, "A dual-prompting for interpretable mental health language models," *ArXiv*, vol. abs/2402.14854, 2024. 21, 26

[113] J. Haltaufderheide and R. Ranisch, "The ethics of chatgpt in medicine and healthcare: a systematic review on large language models (llms)," *NPJ Digital Medicine*, vol. 7, 2024. 21

[114] Y. K. Lee, J. Suh, H. Zhan, J. J. Li, and D. C. Ong, "Large language models produce responses perceived to be empathic," *ArXiv*, vol. abs/2403.18148, 2024. 22

[115] J. Nie, H. Shao, Y. Fan, Q. Shao, H. You, M. Preindl, and X. Jiang, "Llm-based conversational ai therapist for daily functioning screening and psychotherapeutic intervention via everyday smart devices," *ArXiv*, vol. abs/2403.10779, 2024. 23

[116] C. Wu, X. Zhang, Y. Zhang, Y. Wang, and W. Xie, "Pmc-llama: Towards building open-source language models for medicine," 2023. 24

[117] P. Liang, T. Liu, A. Cai, M. Muszynski, R. Ishii, N. Allen, R. Auerbach, D. Brent, R. Salakhutdinov, and L.-P. Morency, "Learning language and multimodal privacy-preserving markers of mood from mobile data," in *Annual Meeting of the Association for Computational Linguistics*, 2021, pp. 4170–4187. 27

[118] Y. Chen, X. Xing, J. Lin, H. Zheng, Z. Wang, Q. Liu, and X. Xu, "Soulchat: Improving llms' empathy, listening, and comfort abilities through fine-tuning with multi-turn empathy conversations," *ArXiv*, vol. abs/2311.00273, 2023. 27

[119] Y.-C. Chang, X. Wang, J. Wang, Y. Wu, K. Zhu, H. Chen, L. Yang, X. Yi, C. Wang, Y. Wang, W. Ye, Y. Zhang, Y. Chang, P. S. Yu, Q. Yang, and X. Xie, "A survey on evaluation of large language models," *ACM Transactions on Intelligent Systems and Technology*, vol. 15, pp. 1 – 45, 2023. 28