PLOS Mental Health

REVIEW

Generative multimodal large language models in mental health care: Applications, opportunities, and challenges

Ariel Soares Teles[1,2]*, **Jaya Chaturvedi**[1], **Tao Wang**[1], **Marcia Scazufca**[3], **Yamiko Msosa**[1], **Daniel Stahl**[1], **Angus Roberts**[1]

1 Department of Biostatistics and Health Informatics, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, United Kingdom, **2** Campus Araioses, Federal Institute of Maranhão, Araioses, Maranhão, Brazil, **3** Hospital das Clínicas HCFMUSP, Faculdade de Medicina, Universidade de São Paulo, São Paulo, Brazil

* ariel.teles@ifma.edu.br,ariel.teles@kcl.ac.uk

Check for updates

OPEN ACCESS

Citation: Teles AS, Chaturvedi J, Wang T, Scazufca M, Msosa Y, Stahl D, et al. (2025) Generative multimodal large language models in mental health care: Applications, opportunities, and challenges. PLOS Ment Health 2(11): e0000488. https://doi.org/10.1371/journal.pmen.0000488

Editor: Karli Montague-Cardoso, PLOS: Public Library of Science, UNITED KINGDOM OF GREAT BRITAIN AND NORTHERN IRELAND

Received: July 29, 2025

Accepted: October 22, 2025

Published: November 6, 2025

Copyright: © 2025 Teles et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: AT is funded by the National Council for Scientific and Technological Development - CNPq (grants 308059/2022-0 and 441817/2023-8).

Abstract

Generative Large Language Models (LLMs) are transforming mental health care by enabling the generation and understanding of human-like text with increasing nuance and contextual awareness. However, mental health is a complex, multidimensional domain that often requires richer sources of information beyond text. This narrative review explores the emerging role of Multimodal LLMs (MLLMs), which are models that integrate diverse input modalities such as speech, images, video, and physiological signals, to incorporate the multifaceted nature of mental states and human interactions. We first outline the foundational principles of MLLMs and their distinction from traditional text-only LLMs. We then synthesize recent empirical studies and experimental applications of MLLMs in mental health research and clinical settings, highlighting their potential to improve diagnostic accuracy, enable real-time monitoring, and support context-aware, personalized interventions. Finally, we outline opportunities for future research and innovation, and discuss key implementation challenges in MLLM-based mental health care.

Introduction

Mental health has become an increasingly critical issue worldwide, with rising rates of anxiety, depression, and other mental health disorders exacerbated by factors such as social isolation, economic pressures, and global crises [1]. The growing recognition of mental health's impact on overall well-being, productivity, and quality of life has underscored the urgent need for effective and accessible care. However, traditional mental health services often face significant barriers, including stigma, high costs, and a shortage of trained professionals, leaving many individuals without adequate support [2]. The mental healthcare sector has already acknowledged the urgency for innovative solutions [3,4]. In response, digital technologies have emerged

TW is supported by the NIHR Maudsley Biomedical Research Centre, Maudsley Charity, King's Together Award and MHaPS Early Career Researcher Seed Funding. AR is funded by the UK Medical Research Council (grant DATAMIND - The Data Hub for Mental Health Informatics Research and Development, number MR/Z504816/1, and grant RE-STAR: Regulating Emotions - Strengthening Adolescent Resilience, number MR/W002493/1). JC is supported with a grant from the Maudsley Charity (grant 2946). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

as a transformative force in addressing these challenges, offering innovative solutions such as mobile applications, teletherapy, chatbots, and virtual reality [5,6]. Such solutions can be further enhanced through the integration of Artificial Intelligence (AI) [7], which enables more responsive, adaptive, and context-aware systems. By harnessing recent advancements in AI, digital technologies hold significant promise for transforming mental health care, expanding access, enhancing personalization, and supporting more effective and scalable interventions.

Large Language Models (LLMs) represent a significant advancement in AI, capable of recognizing, generating, and interacting with human language [8]. These models are trained on vast datasets, enabling them to perform a wide range of tasks, from answering questions and summarizing text to generating creative content and providing personalized recommendations. In healthcare, LLMs have already demonstrated utility [9,10], and hold potential to revolutionize various aspects of care delivery, including diagnostics, patient communication, and clinical decision support [11,12]. For example, they have been investigated as virtual therapists [13] and employed to develop diagnostic and prognostic models based on Electronic Health Records (EHR) data [14]. Therefore, LLMs offer a promising avenue for improving healthcare accessibility, efficiency, and personalization, particularly in areas like mental health care where scalable, personalized, and timely interventions are extremely important.

Developed from 2022 [15], Multimodal LLMs (MLLMs) extend the capabilities of text-only LLMs by integrating and processing more than one data modality, including not only text but also audio, image and video, enabling a more comprehensive understanding of complex inputs [15]. These models are designed to both interpret complex, multimodal inputs and produce outputs in one or more modalities, supporting richer and more flexible interactions. Considerable efforts have been dedicated to creating LLMs that handle multimodal inputs and tasks, ultimately leading to the development of MLLMs [16], which analyse diverse health data to enhance, for example, diagnostics, treatment planning, and personalized care [17]. Therefore, MLLMs represent a significant advancement in healthcare, addressing the multimodal nature of medicine, where handling multiple data types is often required [18,19].

While recent reviews have comprehensively examined the applications, limitations, and ethical implications of text-based LLMs in mental health care and psychiatry [20–29], they have largely focused on unimodal models. These secondary studies are highly relevant and timely, shedding light on key themes such as early screening, conversational agents, and clinical integration. However, none have specifically addressed the emerging class of MLLMs. Our investigation of the existing literature suggests that this is the first review to concentrate on the application of MLLMs in the field of mental health. By examining not only current applications, but also presenting potential areas for future studies, our review contributes a novel perspective on how MLLMs may advance digital mental health solutions.

Although language is a fundamental modality, because human communication is intrinsically linguistic, relying solely on text risks overlooking crucial dimensions of

mental health. By integrating additional modalities, MLLMs can offer a more comprehensive and nuanced understanding of mental health. Therefore, the objective of this narrative review is to explore the emerging role of generative MLLMs in mental health care. Specifically, we synthesize recent empirical and conceptual developments on how MLLMs have been integrated and applied within mental health research and clinical contexts, recognizing that most current evidence remains exploratory or proof-of-concept in nature. We begin by outlining the foundational concepts underlying MLLMs and how they differ from text-only LLMs. We then examine empirical studies and experimental systems that demonstrate their use in mental health settings. Finally, we identify opportunities for future applications and derived studies, as well as important challenges associated with model implementation.

What are generative multimodal LLMs?

A language model is a probabilistic model designed to estimate the likelihood of a sequence of tokens, typically words or subwords, within a given context. Contemporary state-of-the-art language models are predominantly based on the transformer architecture [30], which leverages self-attention mechanisms to capture contextual relationships between tokens across an entire sequence. Transformer-based architectures can be broadly categorized into three classes [31] (Table 1): encoder-only, decoder-only, and encoder-decoder models, each optimized for different types of tasks. Among them, decoder-only and encoder-decoder models are typically considered generative. Decoder-only models are explicitly trained for autoregressive language modelling, where the objective is to predict the next token in a sequence given the preceding tokens. Encoder-decoder models are also generative, particularly in tasks that require transforming an input sequence into a new output sequence.

MLLMs are deep learning models that extend the capabilities of text-only language models by incorporating and jointly processing multiple types of input modalities, such as text, images, audio, video, and physiological signals [15]. While unimodal models operate solely on textual data, MLLMs are designed to integrate heterogeneous information streams to build a richer and more holistic representation of the world [44]. This integration enables the models to perform tasks that require reasoning across modalities [45,46]. For example, generating descriptive text from an image, generating a video from a textual description, or interpreting emotional states to some extent from both spoken language and facial expressions.

MLLMs can be understood as layered architectures that extend unimodal LLMs to process heterogeneous data. Conceptually, they can be viewed as layered systems comprising three core modules (Fig 1): a modality encoder, a modality interface (connector), and a language model backbone, with an optional generator for multimodal output synthesis. At the input stage, modality-specific encoders [47] transform raw signals such as images [48], audio [49], video, or physiological data into vector embeddings [50]. These encoders are typically pre-trained models aligned across modalities. They function as the perceptual components of the system, transforming raw multimodal inputs into structured embeddings that effectively capture and preserve semantic information.

Table 1. Overview of transformer-based language models.

Model Type	Architecture	Typical Input	Output	G	Example Models	M
Encoder-only	Transformer	Text	Text (embedding)	No	BERT [32], RoBERTa [33]	No
Decoder-only	Transformer	Text (prompt)	Text (generated)	Yes	GPT-2, GPT-3 [8]	No
Encoder-decoder	Transformer	Text	Text	Yes	T5 [34], BART [35]	No
MLLM	Hybrid (transformer-based)	Text, images, audio, video	Text, image, audio, video	Yes	LLaVA [36], LLaVA-NeXT-Interleave [37], LLaVA-Video [38], GPT-4o [39,40], GPT-5 [41], Gemini [42], Veo 3 [43]	Yes

Note: G = Generative; M = Multimodal

https://doi.org/10.1371/journal.pmen.0000488.t001

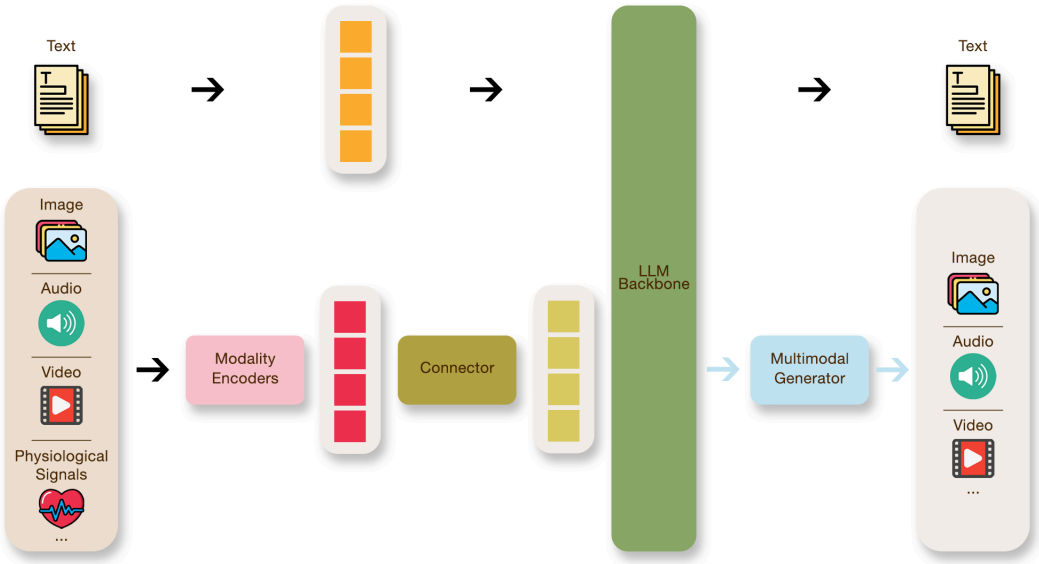


Fig 1. Generic architecture of a Multimodal Large Language Model (MLLM), illustrating how modality-specific encoders, a connector, and an LLM backbone jointly process heterogeneous inputs to generate textual or multimodal outputs.

https://doi.org/10.1371/journal.pmen.0000488.g001

Embeddings are then mapped into the token space of the language model through a connector [51], which aligns different modalities with a unified representation space [15,52]. Depending on the fusion granularity, connectors can operate at the token level or at the feature level. This interface ensures that heterogeneous embeddings are projected into a shared latent space that the LLM backbone can understand. By aligning multimodal inputs in this unified representational space, MLLMs enable seamless joint reasoning across different information types. Once aligned, the LLM backbone performs integrated reasoning by attending jointly to textual tokens and projected multimodal embeddings. The fused representation allows the LLM to produce coherent, context-aware responses that integrate insights from all modalities involved. Finally, some architectures attach a multimodal generator to extend output capabilities beyond text. These generators can produce images, audio, or videos, often via diffusion-based or autoregressive decoders [15].

Some MLLMs are primarily designed to generate textual outputs from non-textual inputs, while others extend this capability toward temporal and multimodal reasoning. For instance, LLaVA [36] interprets visual content through a language interface, producing detailed captions and visual question-answering outputs. Building on this foundation, LLaVA-Video [38] introduces instruction-tuning for dynamic video understanding, enabling fine-grained captioning and reasoning over temporally rich video sequences. Similarly, Veo 3 [43] represents a new generation of generative video models that demonstrate zero-shot visual reasoning across perception, modelling, manipulation, and temporal inference tasks. Trained on web-scale video data with a generative objective, Veo 3 can segment, edit, and reason about physical interactions, solve visual puzzles such as mazes and symmetry completion, and exhibit “chain-of-frames” reasoning analogous to chain-of-thought processes in LLMs [53]. In parallel, GPT-4o [39,40] and GPT-5 [41] exemplify a natively multimodal architecture capable of bidirectional generation across text, image, and audio. They can not only generate descriptive text from visual and auditory inputs but also perform tasks like answering spoken questions, analysing images in real time, and reasoning jointly across text, vision, and audio modalities, demonstrating fluency and flexibility in multimodal interaction.

A key distinction lies between “true” MLLMs and unimodal LLMs integrated within multimodal pipelines. True MLLMs are architecturally designed to natively process and generate across multiple modalities through joint embeddings and

cross-modal reasoning. In contrast, some systems employ unimodal LLMs as downstream reasoning engines in larger multimodal frameworks. They combine multimodal data streams but pass only structured or textualized representations to a unimodal LLM, which then generates textual outputs. While these pipelines do not constitute MLLMs *per se*, they are important to consider in this review because the data itself is multimodal, and unimodal LLMs act as reasoning engines that translate heterogeneous signals into clinically meaningful, natural-language outputs. Moreover, such systems illustrate the transitional stage in which unimodal LLMs extend into multimodal domains, highlighting both the opportunities and the limitations that motivate the development of true MLLMs. In this review, we consistently use the abbreviation "MLLM" for models with native multimodal architectures, while referring separately to unimodal LLMs in multimodal frameworks.

MLLMs applications in mental health

MLLMs hold promise across a wide range of mental health applications [54,55]. This section focuses on recent empirical studies and experimental systems that have already leveraged MLLMs in mental health contexts.

Therapy assistance

Generative LLMs are being explored as tools to assist psychotherapy, both by interacting directly with individuals as virtual therapists and by enhancing the education and training of mental health professionals. For clinicians and trainees, LLMs can serve as accessible educational aids, simulating therapeutic dialogues [56], and offering feedback on clinical communication [57,58]. Simultaneously, LLMs can engage directly with users (i.e., chatbot-based interventions [59]) by delivering responses that emulate therapeutic dialogue [60]. While multimodal capabilities may eventually broaden these applications, current evidence suggests that even text-only models are useful.

Even in the absence of multimodal inputs, LLMs have demonstrated the ability to generate responses that are perceived as empathic, supportive, and aligned with core therapeutic principles. For instance, recent evaluations have shown that licensed clinicians rate AI-generated psychological advice as comparable to (or in some dimensions even more favourable than) expert-authored responses, particularly in terms of emotional and motivational empathy [13,61]. These findings suggest that even without access to multimodal information (e.g., tone of voice, facial expressions), text-based LLMs can simulate therapist-like interactions and generate emotionally resonant messages that, in some evaluations, were rated comparably to those written by professionals.

Recent advances in MLLMs have begun to address the limitations of purely text-based therapeutic systems by integrating visual cues, such as facial expressions and body posture, into the therapeutic dialogue. Zhu et al. [62] proposed a framework that enhances mental health support by embedding a unimodal LLM within a multimodal pipeline. The system combines facial expression analysis, dialogue history, commonsense reasoning, and context-specific counselling strategies. Emotional cues are first extracted from visual signals and transformed into textual descriptions, which are then passed to the LLM for response generation. Although the language model itself remains unimodal, the overall framework integrates multimodal information at the system level. Evaluations showed that this pipeline produced responses rated higher in coherence, empathy, and helpfulness compared to text-only baselines, highlighting the value of using multimodal data even when the core model is unimodal.

Wang et al. [63] conducted a mixed-methods investigation to explore how MLLMs might be meaningfully integrated into psychotherapy for college students, with input from both therapists and student clients. Their findings suggest that, at this stage, MLLMs are best positioned as auxiliary tools that support, rather than replace, human therapists. Participants across all stakeholder groups highlighted potential applications in structured phases of therapy (e.g., initial information gathering, triage matching, and feedback analysis), where MLLMs could assist by analysing multimodal inputs (e.g., speech, facial expressions, self-report data) and generating summaries or suggestions. Both therapists and students also expressed interest in real-time emotion recognition and personalized support, though concerns about emotional depth, ethical risks, and privacy remain strong.

Mental health monitoring

A key application of MLLMs is their potential to assist in the early detection and intervention of mental health conditions. These capabilities emerge from their integration into monitoring technologies (e.g., personal sensing [64], digital phenotyping [65]). By analysing data collected from smartphones, wearables, and other ambient sensors (e.g., speech patterns, typing behaviour, facial expressions), LLMs can help identify subtle and context-specific changes that may signal emerging psychological distress. When deployed in conjunction with these sensing systems, LLMs can interpret complex, multimodal behavioural data to flag at-risk individuals, deliver timely feedback, or trigger human-led follow-up. This enables scalable, continuous, and personalized mental health monitoring that moves toward just-in-time adaptive interventions [66].

A few studies have already initiated the integration of unimodal LLMs within multimodal monitoring frameworks. One example of such integration is SereniSens [67], a multimodal AI framework designed to monitor and predict stress levels during sleep by combining physiological sensing and language-based interaction. The system uses biosignals such as heart rate variability, respiration rate, blood oxygen level, and body temperature to classify users into five stress levels, using conventional machine learning algorithms. While the input data is multimodal, the LLM component is unimodal: it receives structured outputs from the predictive models and transforms them into personalized, therapeutic messages. A fine-tuned version of GPT-3.5 Turbo is used to interpret the predicted stress levels and associated metrics, generating context-sensitive dialogue through a chatbot interface. While not a true MLLM, this architecture exemplifies how LLMs can function as natural language reasoning engines in larger sensor-based systems, bridging the gap between low-level physiological signal analysis and high-level, user-facing mental health support. Rather than directly ingesting multimodal inputs, the LLM here operates downstream, making it a powerful tool for translating quantitative sensor data into accessible, empathic guidance.

A similar architecture is employed in the Intelligent Phenotype Analysis Suite (IPAS) [68], a tool designed to streamline the analysis of digital phenotyping data. IPAS supports researchers in extracting and interpreting behavioural and physiological signals (e.g., GPS, accelerometer, text message logs) captured through platforms like Beiwe [69] and RADAR-Base [70]. As in the SereniSens framework, the LLM used in IPAS is not multimodal. The system utilizes program-aided prompting techniques to generate descriptions, visualizations, and predictions based on time-series data, while also enabling natural language explanations of findings. This allows researchers, including those with limited programming experience, to engage with complex behavioural datasets via a chatbot-style interface. IPAS exemplifies a growing class of applications in which unimodal LLMs function as reasoning and communication layers within multimodal pipelines, in which multimodal data are textualized [14].

Although originally developed for hypertension management, the Hyper-DREAM platform [71] offers valuable insights into how MLLMs can support mental health monitoring when integrated with digital phenotyping systems. The platform collects rich multimodal data (e.g., self-reported health information, behavioural patterns, and physiological signals like blood pressure and heart rate), and synthesizes this information into individualized digital phenotypes. A fine-tuned LLM, deployed as a chatbot, interprets these data streams to generate personalized feedback, health education, and emotional support. While the primary focus is cardiovascular care, the system incorporates mental health indicators such as sleep quality, stress, and mood, demonstrating the broader potential of LLMs to process passively collected multimodal data.

Diagnostic support

Multimodal LLMs can assist in the diagnostic process by analysing diverse data sources alongside clinical text to identify signs of mental health conditions, including complex or ambiguous cases [72,73]. Their ability to integrate heterogeneous data makes them valuable tools for supporting clinicians in early detection and differential diagnosis.

Toto et al. [74] introduced AudiBERT, a deep transfer learning framework for depression screening that integrates both textual and audio data from clinical interviews. By combining pre-trained representations from BERT (for textual

transcripts), an encoder-only LLM [32], and models such as Wav2Vec and SincNet (for vocal signals), AudiBERT captures both verbal and non-verbal indicators of depression. A dual self-attention mechanism fuses these modalities into a shared representation optimized for classification. Similarly, Ali et al. [75] investigated the use of MLLMs for detecting depression and Post Traumatic Stress Disorder (PTSD) through both text and audio data from clinical interviews. Evaluating models like Gemini 1.5 Pro and GPT-4o mini, the authors demonstrate that combining audio and text inputs improves diagnostic performance compared to unimodal approaches. These works illustrate the potential of multimodal solutions to leverage the rich, multimodal nature of human communication for accurate and scalable diagnostic support.

A recent study by Sadeghi et al. [76] demonstrated how LLMs can be combined with facial expression data to enhance depression detection using a fully automated, multimodal approach. The proposed system leverages GPT-based models to extract depression-related features from interview transcripts and combines them with visual features such as facial action units and eye gaze derived from video data. By training a regression model to predict PHQ-8 scores, the authors showed that the multimodal approach outperformed text- or video-only baselines on a held-out test set, particularly when supported by speech quality assessment. Notably, text-based features had the highest predictive power, but visual cues provided complementary information that improved robustness, especially under conditions with reduced verbal expressivity. This work highlights the potential of MLLMs to contribute to objective, scalable, and clinically relevant assessments of depression severity, including in settings where access to trained professionals may be limited.

Empirical evidence landscape

As observed throughout the previous sections, and consistent with findings from other reviews [16,77], the empirical literature on MLLMs in mental health remains in an early stage of development. Most available studies are proof-of-concept or mixed-methods evaluations focused on technical feasibility, user experience, or therapeutic potential rather than validated clinical outcomes. A few works have included structured human evaluations, but large-scale randomized or longitudinal clinical studies are yet to emerge. The current evidence base therefore reflects an exploratory phase of research, where multimodal systems are primarily assessed in simulated or pilot settings. This limited empirical maturity underscores both the novelty of the field and the need for rigorous, clinically validated studies to evaluate effectiveness and safety.

Opportunities

While current applications of MLLMs in mental health are still in the early stages, they point to a wide range of promising future directions. This section outlines emerging research opportunities where MLLMs could meaningfully expand the reach, effectiveness, and personalization of mental health care.

Therapist–AI collaboration

A promising opportunity for MLLMs lies in their potential to support therapists [78]. Rather than functioning autonomously, these models can act as reflective partners in post-session analysis. To illustrate this potential, consider a hypothetical scenario in which a psychotherapist providing telehealth care to a patient with chronic emotional dysregulation and interpersonal difficulties begins to notice limited therapeutic progress. To enhance their clinical insight, the therapist discusses with the patient the possibility of involving an AI model to assist in reviewing recorded sessions. With informed consent, the MLLM is granted access to video, audio, and transcript data from therapy sessions, enabling it to generate structured, therapist-facing reports. The model's output supports the therapist by summarizing the session's content, identifying which interventions were effective or ineffective, and suggesting targeted actions for future sessions. Rather than replacing the therapist's judgment, the MLLM serves as a collaborative analytic tool—highlighting patterns across sessions, surfacing overlooked cues, and offering clinical suggestions grounded in best practices. An illustrative post-session report might include the following elements:

- Session summary: key topics discussed, emotional tone, and therapeutic strategies employed;
- Helpful elements: Moments where interventions (e.g., reflective listening, use of silence) elicited engagement or emotional expression;
- Less effective elements: instances where techniques (e.g., cognitive reframing, goal-setting) were met with resistance or disengagement;
- Recommendations: suggestions for revisiting specific themes, adjusting timing or delivery of interventions, or attending to nonverbal cues in subsequent sessions;
- Longitudinal patterns: emerging dynamics across multiple sessions, such as recurring ruptures or points of progress.

This example illustrates how MLLMs can act as ethically integrated collaborators, validated by the patient and accountable to the therapist's professional judgment. By supporting longitudinal tracking and offering fresh perspectives on therapeutic interactions, these models have the potential to enhance care delivery in complex or stagnant cases, particularly in remote or digitally mediated settings. However, the actual impact of such collaboration remains an open research question. The integration of MLLMs into therapeutic practice represents a promising area for future research, particularly studies that examine whether this collaboration yields measurable clinical benefits and how it is perceived by both patients [79,80] and mental health professionals [81]. Understanding these perspectives is essential to ensure that MLLM-assisted care is not only technically feasible, but also trusted, acceptable, and aligned with ethical standards in real-world practice.

Supporting personalized interventions

While innovative solutions have already leveraged text-only LLMs for mental health support [20,21,27,82], MLLMs open new frontiers by incorporating diverse data sources. These models hold considerable promise for advancing the next generation of digital mental health interventions [6]. MLLMs can enhance traditional care models by enabling early and low-threshold access to support. By continuously analysing multimodal data streams (e.g., written reflections, mood logs, facial expressions, or biometric signals from wearables), these models can identify subtle shifts in emotional or behavioural states. For example, a young adult using a self-guided mental health app might regularly submit short journal entries and selfies. Over time, the MLLM could detect signs of distress (e.g., negative affect, social withdrawal, disorganized appearance) and respond with tailored, just-in-time interventions [66], such as grounding techniques, cognitive reframing prompts, or a gentle recommendation to contact a support person. In this way, MLLMs function not only as personalized companions but also as early-warning systems, offering scalable and cost-effective support that bridges gaps in traditional care.

Education and awareness

Education and awareness are critical components of mental health promotion, as they influence how individuals recognize symptoms, seek support, and respond to others experiencing distress. Yet, traditional educational approaches might fail to reach all audiences effectively, especially those with limited health literacy, learning differences, or cultural and linguistic barriers. MLLMs can offer a transformative approach by delivering mental health information through a combination of text, images, audio, and video. This multimodal strategy aligns with how people naturally process and retain information, improving comprehension, engagement, and recall [83–85]. For instance, an MLLM could convert complex content about depression or anxiety into a short, accessible explainer video using a text-to-video generation model [86] (e.g., Veo 3 [43], LLaVA-Video [38]), allowing users to engage with the material in a more intuitive and emotionally resonant way. These adaptive capabilities are especially useful for supporting individuals with different cognitive profiles, sensory impairments, or educational needs. By tailoring information delivery to the learner, MLLMs can help reduce stigma, foster early identification of mental health concerns, and encourage timely help-seeking, ultimately contributing to more informed, empowered, and mentally healthy communities.

Developing MLLM-based predictive models

A particularly promising application of MLLMs in mental health care is the development of predictive models that leverage multimodal data. While prediction models based on structured or unstructured data alone have shown value, the integration of both modalities using multimodal machine/deep learning has demonstrated superior performance in prediction tasks [87]. Applying fusion strategies (e.g., early fusion, intermediate fusion, late fusion) [88] offers flexible pathways for building robust predictive tools. These strategies enable MLLMs to model cross-modal relationships [89] that are essential in mental health scenarios, such as detecting early signs of relapse, forecasting crisis events, or identifying non-adherence to therapy or treatment. Each fusion strategy offers unique strengths: early fusion supports joint learning from the outset, intermediate fusion preserves modality-specific features before integration, and late fusion combines independent predictions; making the choice of strategy dependent on data characteristics and clinical goals [14]. These capabilities open a promising direction for future work: designing and validating MLLM-based predictive systems tailored to mental health care.

Specializing MLLMs for mental health tasks

Another important opportunity for expanding the clinical utility of MLLMs lies in the ability to adapt them to specific mental health tasks through fine-tuning. To enhance their performance on specific tasks or within particular domains, both unimodal and multimodal language models can undergo fine-tuning [90]. This is a process in which a pre-trained model is further trained on a smaller, task-specific dataset. While large-scale pre-training equips models with broad linguistic and perceptual capabilities, fine-tuning allows for specialization by adapting the model's parameters to the statistical characteristics and objectives of the target application. In some cases, smaller LLMs that have been carefully fine-tuned on domain-relevant data can outperform much larger general-purpose models [91], particularly when computational efficiency, data privacy, or deployment constraints are important considerations. For example, a lightweight model fine-tuned on therapy transcripts and synchronized voice tone or facial expression data might produce more contextually appropriate and emotionally attuned responses in a mental health setting than a general-purpose model that lacks exposure to such multimodal interactions [74,75]. Fine-tuning thus serves not only to improve performance but also to support the development of specialized, resource-efficient models tailored for high-stakes applications such as mental health care.

Challenges

MLLMs offer promising opportunities for advancing mental health care, but their adoption also raises critical challenges that must be carefully addressed. In this section, we highlight current key concerns.

Accuracy and reliability

In multimodal mental health modelling, accuracy depends not only on language understanding but also on the correct fusion and temporal synchronization of heterogeneous signals. Small errors in temporal alignment between speech, facial cues, and physiological responses can yield spurious correlations or misattributed emotional states. Moreover, modality-specific noise (e.g., lighting variation in video, background noise in audio, motion artifacts in wearables) propagates through embedding layers and can distort downstream reasoning. Ensuring the accuracy and reliability of MLLMs is therefore a critical prerequisite for their safe integration into mental health care. While these models show promise in generating plausible and empathetic responses, their clinical performance must be rigorously validated through systematic testing, benchmarking, and real-world trials. Without such validation, there is a risk that models may produce outputs that are factually incorrect, contextually inappropriate, or potentially harmful hallucinations [92]. To address this, recent studies have begun evaluating the performance of LLMs and MLLMs in psychological tasks under experimental conditions [13,61]. These efforts represent important first steps, but broader, more diverse, and clinically grounded assessments are needed to establish trust and ensure safe deployment in high-stakes mental health contexts.

Bias and fairness

Biased interpretations of multimodal signals may reinforce stereotypes or misclassify distress among underrepresented populations. MLLMs raise serious concerns about bias and fairness [93]. These models are trained on large-scale multimodal datasets that often encode existing societal and cultural biases, including those related to race, gender, age, and mental health stigma [94–96]. Textual data may reflect biased language in how mental health conditions are described, while visual data may underrepresent marginalized groups, resulting in uneven model performance across populations. In multimodal systems, these biases can be compounded, further amplifying disparities. For example, biased models may produce higher-quality outputs for certain demographic groups while misrepresenting or oversimplifying the mental health experiences of others [97]. This not only risks unequal access to support but may also reinforce harmful stereotypes. Addressing these issues requires deliberate efforts to curate diverse and representative training data, incorporate input from a wide range of stakeholders, and ensure transparency about model limitations [98]. Without these safeguards, MLLMs risk perpetuating the very inequities they aim to alleviate.

User trust and acceptance

The successful integration of MLLMs in mental health care depends heavily on user trust and acceptance, both from clinicians and patients. Mental health professionals need to feel confident that these tools support, rather than undermine, their clinical judgment, while patients must trust the system enough to engage meaningfully and disclose sensitive information. Building this trust requires not only high-quality and ethical model behaviour, but also interpretability and explainability. However, MLLMs, which are based on deep learning, often behave as opaque “black boxes”, making it difficult to trace which modality or temporal window drove an output (e.g., a prediction or a therapeutic suggestion). This opacity undermines confidence in automated recommendations. When users can understand how a model arrives at its outputs, they are more likely to evaluate its reliability and limitations appropriately. Recent studies have shown that incorporating explainability mechanisms can positively influence trust in mental health applications of AI [99]. As MLLMs continue to evolve, ensuring transparent, interpretable interactions will be critical for fostering informed trust and sustained adoption.

Regulatory and ethical issues

The use of MLLMs in mental health care introduces complex regulatory and ethical challenges that current oversight frameworks are not fully equipped to address. MLLMs, by processing and generating outputs across multiple data types, raise unique concerns around data protection, transparency, and accountability. Yet, as noted in recent policy reviews [100], these systems often operate outside existing regulatory classifications, such as those governing medical devices, which creates uncertainty regarding their legal status and the responsibilities of developers and clinicians. Ethical evaluations have further highlighted critical risks, including the potential for harm due to misinformation, breaches of privacy, opaque decision-making, and the erosion of human autonomy in therapeutic contexts [4]. As such, responsible deployment of MLLMs requires not only alignment with evolving regulations but also proactive ethical design practices, including transparent communication of model limitations, rigorous testing, and meaningful involvement of diverse stakeholders [101].

Deployment and interoperability

Integrating MLLMs into mental health workflows presents deployment barriers because mental health services may be delivered in resource-constrained or community-based settings rather than highly digitized environments. Deployment of MLLMs in clinical environments demands rapid inference on constrained hardware (e.g., mobile devices, legacy workstations, point-of-care kiosks), yet leading models often require substantial computational resources, precluding use in low-resource settings [102]. Employing model compression methods (e.g., parameter quantization, model pruning, low-rank decomposition, knowledge distillation) [103] and fine-tuning model variants can mitigate latency spikes and reduce

computational load. Also, like other clinical decision-support tools, MLLMs must integrate seamlessly with existing EHR systems and clinical workflows to maximize their utility [14]. Achieving this integration demands close collaboration among clinicians, researchers, and health-informatics specialists [104].

Conclusion

In this review, we aimed to equip researchers, clinicians, and developers with a comprehensive understanding of the current capabilities, limitations, and future directions for the deployment of MLLMs in mental health contexts. We synthesized emerging empirical and conceptual evidence on the use of generative MLLMs in mental health care, highlighting their potential to enhance diagnostic support, therapy, monitoring, and education through richer, context-aware, and personalized interactions. While existing applications demonstrate promising technical feasibility and therapeutic potential, the empirical foundation of this field remains preliminary, and critical challenges persist, particularly regarding clinical validation, fairness, user trust, and regulatory oversight. Addressing these challenges will be essential to ensure that MLLMs evolve into safe, effective, and ethically aligned tools. Ultimately, with continued research and responsible development, MLLMs have the potential to augment mental health care delivery in ways that are both human-centred and technologically innovative.

Author contributions

Conceptualization: Ariel Soares Teles, Jaya Chaturvedi, Tao Wang, Marcia Scazufca, Yamiko Msosa, Daniel Stahl, Angus Roberts.

Formal analysis: Ariel Soares Teles, Jaya Chaturvedi, Tao Wang, Marcia Scazufca, Yamiko Msosa, Daniel Stahl, Angus Roberts.

Investigation: Ariel Soares Teles.

Writing – original draft: Ariel Soares Teles, Jaya Chaturvedi, Tao Wang, Marcia Scazufca, Yamiko Msosa.

Writing – review & editing: Ariel Soares Teles, Jaya Chaturvedi, Tao Wang, Marcia Scazufca, Yamiko Msosa, Daniel Stahl, Angus Roberts.

References

1. World Health Organization. World mental health report: transforming mental health for all. 2022. https://www.who.int/publications/i/item/9789240049338
2. Thornicroft G, Chatterji S, Evans-Lacko S, Gruber M, Sampson N, Aguilar-Gaxiola S, et al. Undertreatment of people with major depressive disorder in 21 countries. Br J Psychiatry. 2017;210(2):119–24. https://doi.org/10.1192/bjp.bp.116.188078 PMID: 27908899
3. D'Alfonso S. AI in mental health. Curr Opin Psychol. 2020;36:112–7. https://doi.org/10.1016/j.copsyc.2020.04.005 PMID: 32604065
4. Vilaza GN, McCashin D. Is the automation of digital mental health ethical? Applying an ethical framework to chatbots for cognitive behaviour therapy. Frontiers in Digital Health. 2021;3.
5. Torous J, Bucci S, Bell IH, Kessing LV, Faurholt-Jepsen M, Whelan P, et al. The growing field of digital psychiatry: current evidence and the future of apps, social media, chatbots, and virtual reality. World Psychiatry. 2021;20(3):318–35. https://doi.org/10.1002/wps.20883 PMID: 34505369
6. Löchner J, Carlbring P, Schuller B, Torous J, Sander LB. Digital interventions in mental health: an overview and future perspectives. Internet Interv. 2025;40:100824. https://doi.org/10.1016/j.invent.2025.100824 PMID: 40330743
7. Cruz-Gonzalez P, He AW-J, Lam EP, Ng IMC, Li MW, Hou R, et al. Artificial intelligence in mental health care: a systematic review of diagnosis, monitoring, and intervention applications. Psychol Med. 2025;55:e18. https://doi.org/10.1017/S0033291724003295 PMID: 39911020
8. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language models are few-shot learners. In: Larochelle H, Ranzato M, Hadsell R, Balcan MF, Lin H, editors. Advances in Neural Information Processing Systems. Curran Associates, Inc. 2020. p. 1877–901.
9. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature. 2023;620(7972):172–80. https://doi.org/10.1038/s41586-023-06291-2 PMID: 37438534

10. Neveditsin N, Lingras P, Mago V. Clinical insights: a comprehensive review of language models in medicine. PLOS Digit Health. 2025;4(5):e0000800. https://doi.org/10.1371/journal.pdig.0000800 PMID: 40338967
11. Wang D, Zhang S. Large language models in medical and healthcare fields: applications, advances, and challenges. Artif Intell Rev. 2024;57(11). https://doi.org/10.1007/s10462-024-10921-0
12. He K, Mao R, Lin Q, Ruan Y, Lan X, Feng M, et al. A survey of large language models for healthcare: from data, technology, and applications to accountability and ethics. Information Fusion. 2025;118:102963. https://doi.org/10.1016/j.inffus.2025.102963
13. Hatch SG, Goodman ZT, Vowels L, Hatch HD, Brown AL, Guttman S, et al. Correction: when ELIZA meets therapists: a turing test for the heart and mind. PLOS Ment Health. 2025;2(8):e0000426. https://doi.org/10.1371/journal.pmen.0000426
14. Teles AS, de Moura IR, Silva F, Roberts A, Stahl D. EHR-based prediction modelling meets multimodal deep learning: a systematic review of structured and textual data fusion methods. Information Fusion. 2025;118:102981. https://doi.org/10.1016/j.inffus.2025.102981
15. Yin S, Fu C, Zhao S, Li K, Sun X, Xu T, et al. A survey on multimodal large language models. Natl Sci Rev. 2024;11(12):nwae403. https://doi.org/10.1093/nsr/nwae403 PMID: 39679213
16. Xiao H, Zhou F, Liu X, Liu T, Li Z, Liu X, et al. A comprehensive survey of large language models and multimodal large language models in medicine. Information Fusion. 2025;117:102888. https://doi.org/10.1016/j.inffus.2024.102888
17. Mesko B. The impact of multimodal large language models on health care's future. J Med Internet Res. 2023;25:e52865.
18. AlSaad R, Abd-alrazaq A, Boughorbel S, Ahmed A, Renault MA, Damseh R. Multimodal large language models in health care: applications, challenges, and future outlook. J Med Internet Res. 2024;26(e59505).
19. Dang T, Jia H. Multimodal large language models in human-centered health: practical insights. IEEE Pervasive Comput. 2024;23(4):87–93. https://doi.org/10.1109/mprv.2024.3499513
20. Guo Z, Lai A, Thygesen JH, Farrington J, Keen T, Li K. Large language models for mental health applications: systematic review. JMIR Ment Health. 2024;11:e57400. https://doi.org/10.2196/57400 PMID: 39423368
21. Liu Z, Bao Y, Zeng S, Qian R, Deng M, Gu A. Large language models in psychiatry: current applications, limitations, and future scope. Big Data Mining and Analytics. 2024;7(4):1148–68.
22. Kolding S, Lundin RM, Hansen L, Østergaard SD. Use of generative artificial intelligence (AI) in psychiatry and mental health care: a systematic review. Acta Neuropsychiatr. 2024;37:e37. https://doi.org/10.1017/neu.2024.50 PMID: 39523628
23. Xian X, Chang A, Xiang Y-T, Liu MT. Debate and dilemmas regarding generative AI in mental health care: scoping review. Interact J Med Res. 2024;13:e53672. https://doi.org/10.2196/53672 PMID: 39133916
24. Volkmer S, Meyer-Lindenberg A, Schwarz E. Large language models in psychiatry: opportunities and challenges. Psychiatry Res. 2024;339:116026. https://doi.org/10.1016/j.psychres.2024.116026 PMID: 38909412
25. Obradovich N, Khalsa SS, Khan W, Suh J, Perlis RH, Ajilore O, et al. Opportunities and risks of large language models in psychiatry. NPP Digit Psychiatry Neurosci. 2024;2(1):8. https://doi.org/10.1038/s44277-024-00010-z PMID: 39554888
26. Yuan A, Garcia Colato E, Pescosolido B, Song H, Samtani S. Improving workplace well-being in modern organizations: a review of large language model-based mental health chatbots. ACM Trans Manage Inf Syst. 2025;16(1):1–26. https://doi.org/10.1145/3701041
27. Jin Y, Liu J, Li P, Wang B, Yan Y, Zhang H. The applications of large language models in mental health: scoping review. J Med Internet Res. 2025;27(e69284).
28. Hua Y, Na H, Li Z, Liu F, Fang X, Clifton D, et al. A scoping review of large language models for generative tasks in mental health care. NPJ Digit Med. 2025;8(1):230. https://doi.org/10.1038/s41746-025-01611-4 PMID: 40307331
29. Yu L, Ge Y, Ansari S, Imran M, Ahmad W. Multimodal sensing-enabled large language models for automated emotional regulation: a review of current technologies, opportunities, and challenges. Sensors. 2025;25(15).
30. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Advances in neural information processing systems. 2017;30.
31. Lin T, Wang Y, Liu X, Qiu X. A survey of transformers. AI Open. 2022;3:111–32. https://doi.org/10.1016/j.aiopen.2022.10.001
32. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. 2019. https://arxiv.org/abs/1810.04805
33. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D. Roberta: a robustly optimized bert pretraining approach. arXiv preprint 2019. https://arxiv.org/abs/1907.11692
34. Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M. Exploring the limits of transfer learning with a unified text-to-text transformer. J Mach Learn Res. 2020;21(1).
35. Lewis M, Liu Y, Goyal N, Ghazvininejad M, Mohamed A, Levy O, et al. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020. https://doi.org/10.18653/v1/2020.acl-main.703
36. Liu H, Li C, Wu Q, Lee YJ. Visual instruction tuning. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. 2023.
37. Li F, Zhang R, Zhang H, Zhang Y, Li B, Li W. LLaVA-NeXT-interleave: tackling multi-image, video, and 3D in large multimodal models. 2024. https://arxiv.org/abs/2407.07895

38. Zhang Y, Wu J, Li W, Li B, Ma Z, Liu Z. LLaVA-video: video instruction tuning with synthetic data. arXiv preprint 2025. https://arxiv.org/abs/2410.02713
39. Hurst A, Lerer A, Goucher AP, Perelman A. OpenAI. GPT-4o system card. arXiv preprint 2024. https://arxiv.org/abs/2410.21276
40. OpenAI. Hello GPT-4o. 2024. https://openai.com/index/hello-gpt-4o/
41. OpenAI. GPT-5 system card. 2025. https://cdn.openai.com/gpt-5-system-card.pdf
42. Team G, Anil R, Borgeaud S, Alayrac JB, Yu J, Soricut R. Gemini: a family of highly capable multimodal models. arXiv preprint2025. https://arxiv.org/abs/2312.11805
43. Wiedemer T, Li Y, Vicol P, Gu SS, Matarese N, Swersky K. Video models are zero-shot learners and reasoners. arXiv preprint2025. https://arxiv.org/abs/2509.20328
44. Huang D, Yan C, Li Q, Peng X. From large language models to large multimodal models: a literature review. Applied Sciences. 2024;14(12):5068. https://doi.org/10.3390/app14125068
45. Team G, Anil R, Borgeaud S, Alayrac JB, Yu J, Soricut R. Gemini: a family of highly capable multimodal models. arXiv preprint2023. https://arxiv.org/abs/2312.11805
46. Wu J, Gan W, Chen Z, Wan S, Yu PS. Multimodal large language models: a survey. In: 2023 IEEE International Conference on Big Data (BigData). 2023. p. 2247–56. https://doi.org/10.1109/bigdata59044.2023.10386743
47. Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. Learning transferable visual models from natural language supervision. In: International conference on machine learning. 2021. p. 8748–63.
48. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T. An image is worth 16x16 words: transformers for image recognition at scale. In: International Conference on Learning Representations. 2021.
49. Schneider S, Baevski A, Collobert R, Auli M. wav2vec: Unsupervised pre-training for speech recognition. arXiv preprint 2019. https://arxiv.org/abs/1904.05862
50. Driess D, Xia F, Sajjadi MSM, Lynch C, Chowdhery A, Ichter B. PaLM-E: an embodied multimodal language model. In: Proceedings of the 40th International Conference on Machine Learning. 2023. http://jmlr.org
51. Lin J, Chen H, Zhu D, Shen X. To preserve or to compress: an in-depth study of connector selection in multimodal large language models. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. 2024. p. 5666–80. https://doi.org/10.18653/v1/2024.emnlp-main.325
52. Wang X, Chen G, Qian G, Gao P, Wei X-Y, Wang Y, et al. Large-scale multi-modal pre-trained models: a comprehensive survey. Mach Intell Res. 2023;20(4):447–82. https://doi.org/10.1007/s11633-022-1410-8
53. Wei J, Wang X, Schuurmans D, Bosma M, Ichter B, Xia F. Chain-of-thought prompting elicits reasoning in large language models. arXiv preprint 2023. https://arxiv.org/abs/2201.11903
54. Demszky D, Yang D, Yeager DS, Bryan CJ, Clapper M, Chandhok S, et al. Using large language models in psychology. Nat Rev Psychol. 2023. https://doi.org/10.1038/s44159-023-00241-5
55. Thapa S, Adhikari S. GPT-4o and multimodal large language models as companions for mental wellbeing. Asian J Psychiatr. 2024;99:104157. https://doi.org/10.1016/j.ajp.2024.104157 PMID: 39053243
56. Yosef S, Zisquit M, Cohen B, Klomek Brunstein A, Bar K, Friedman D. Assessing motivational interviewing sessions with AI-generated patient simulations. In: Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024). 2024. p. 1–11.
57. Louie R, Nandi A, Fang W, Chang C, Brunskill E, Yang D. Roleplay-doh: enabling domain-experts to create LLM-simulated patients via eliciting and adhering to principles. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. 2024. p. 10570–603. https://doi.org/10.18653/v1/2024.emnlp-main.591
58. Wang R, Milani S, Chiu JC, Zhi J, Eack SM, Labrum T, et al. PATIENT-ψ: using large language models to simulate patients for training mental health professionals. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. 2024. p. 12772–97. https://doi.org/10.18653/v1/2024.emnlp-main.711
59. Limpanopparat S, Gibson E, Harris DA. User engagement, attitudes, and the effectiveness of chatbots as a mental health intervention: a systematic review. Computers in Human Behavior: Artificial Humans. 2024;2(2):100081. https://doi.org/10.1016/j.chbah.2024.100081
60. Siddals S, Torous J, Coxon A. "It happened to be the perfect thing": experiences of generative AI chatbots for mental health. Npj Ment Health Res. 2024;3(1):48. https://doi.org/10.1038/s44184-024-00097-4 PMID: 39465310
61. Franke Föyen L, Zapel E, Lekander M, Hedman-Lagerlöf E, Lindsäter E. Artificial intelligence vs. human expert: licensed mental health clinicians' blinded evaluation of AI-generated and expert psychological advice on quality, empathy, and perceived authorship. Internet Interv. 2025;41:100841. https://doi.org/10.1016/j.invent.2025.100841 PMID: 40525210
62. Zhu Z, Wang S, Wang Y, Wu J. Integrating visual modalities with large language models for mental health support. In: Proceedings of the 31st International Conference on Computational Linguistics. 2025. p. 8939–54.
63. Wang J, Sheng Y, He Q, Hu H, Liu S, Gu F. Enhancing psychotherapeutic alliance in college: when and how to integrate multimodal large language models in psychotherapy. arXiv preprint 2025. https://arxiv.org/abs/2502.00229
64. Mohr DC, Zhang M, Schueller SM. Personal sensing: understanding mental health using ubiquitous sensors and machine learning. Annu Rev Clin Psychol. 2017;13:23–47. https://doi.org/10.1146/annurev-clinpsy-032816-044949 PMID: 28375728

65. Mendes JPM, Moura IR, Van de Ven P, Viana D, Silva FJS, Coutinho LR, et al. Sensing apps and public data sets for digital phenotyping of mental health: systematic review. J Med Internet Res. 2022;24(2):e28735. https://doi.org/10.2196/28735 PMID: 35175202

66. van Genugten CR, Thong MSY, van Ballegooijen W, Kleiboer AM, Spruijt-Metz D, Smit AC, et al. Beyond the current state of just-in-time adaptive interventions in mental health: a qualitative systematic review. Front Digit Health. 2025;7:1460167. https://doi.org/10.3389/fdgth.2025.1460167 PMID: 39935463

67. Wahab O, Adda M, Zrira N. SereniSens: a multimodal AI framework with LLMs for stress prediction through sleep biometrics. Procedia Computer Science. 2024;251:342–9. https://doi.org/10.1016/j.procs.2024.11.119

68. Nissen D, Yu T, Babtista R, Shang Y. Utilizing Large Language Models (LLMs) in data analysis pipeline for digital phenotyping: description, prediction, and visualization. In: 2024 IEEE International Conference on Big Data (BigData). 2024. p. 7513–20. https://doi.org/10.1109/bigdata62323.2024.10825517

69. Onnela J-P, Dixon C, Griffin K, Jaenicke T, Minowada L, Esterkin S, et al. Beiwe: a data collection platform for high-throughput digital phenotyping. JOSS. 2021;6(68):3417. https://doi.org/10.21105/joss.03417

70. Ranjan Y, Rashid Z, Stewart C, Conde P, Begale M, Verbeeck D, et al. RADAR-Base: open source mobile health platform for collecting, monitoring, and analyzing data using sensors, wearables, and mobile devices. JMIR Mhealth Uhealth. 2019;7(8):e11734. https://doi.org/10.2196/11734 PMID: 31373275

71. Wang Y, Zhu T, Zhou T, Wu B, Tan W, Ma K, et al. Hyper-DREAM, a multimodal digital transformation hypertension management platform integrating large language model and digital phenotyping: multicenter development and initial validation study. J Med Syst. 2025;49(1):42. https://doi.org/10.1007/s10916-025-02176-1 PMID: 40172683

72. Ríos-Hoyo A, Shan NL, Li A, Pearson AT, Pusztai L, Howard FM. Evaluation of large language models as a diagnostic aid for complex medical cases. Frontiers in Medicine. 2024;11.

73. Gargari OK, Fatehi F, Mohammadi I, Firouzabadi SR, Shafiee A, Habibi G. Diagnostic accuracy of large language models in psychiatry. Asian J Psychiatr. 2024;100:104168. https://doi.org/10.1016/j.ajp.2024.104168 PMID: 39111087

74. Toto E, Tlachac M, Rundensteiner EA. AudiBERT: a deep transfer learning multimodal classification framework for depression screening. In: Proceedings of the 30th ACM International Conference on Information & Knowledge Management. 2021. p. 4145–54.

75. Ali AA, Fouda AE, Hanafy RJ, Fouda ME. Leveraging audio and text modalities in mental health: a study of LLMs performance. arXiv preprint 2024. https://arxiv.org/abs/2412.10417

76. Sadeghi M, Richer R, Egger B, Schindler-Gmelch L, Rupp LH, Rahimi F, et al. Harnessing multimodal approaches for depression detection using large language models and facial expressions. Npj Ment Health Res. 2024;3(1):66. https://doi.org/10.1038/s44184-024-00112-8 PMID: 39715786

77. Gupta P, Zhang Z, Song M, Michalowski M, Hu X, Stiglic G, et al. Rapid review: growing usage of multimodal large language models in healthcare. J Biomed Inform. 2025;169:104875. https://doi.org/10.1016/j.jbi.2025.104875 PMID: 40754135

78. Sezgin E, McKay I. Behavioral health and generative AI: a perspective on future of therapies and patient care. Npj Ment Health Res. 2024;3(1):25. https://doi.org/10.1038/s44184-024-00067-w PMID: 38849499

79. Benda N, Desai P, Reza Z, Zheng A, Kumar S, Harkins S, et al. Patient perspectives on AI for mental health care: cross-sectional survey study. JMIR Ment Health. 2024;11:e58462. https://doi.org/10.2196/58462 PMID: 39293056

80. Reading Turchioe M, Desai P, Harkins S, Kim J, Kumar S, Zhang Y, et al. Differing perspectives on artificial intelligence in mental healthcare among patients: a cross-sectional survey study. Front Digit Health. 2024;6:1410758. https://doi.org/10.3389/fdgth.2024.1410758 PMID: 39679142

81. Sahoo RK, Sahoo KC, Negi S, Baliarsingh SK, Panda B, Pati S. Health professionals' perspectives on the use of Artificial Intelligence in healthcare: a systematic review. Patient Educ Couns. 2025;134:108680. https://doi.org/10.1016/j.pec.2025.108680 PMID: 39893988

82. Diniz EJS, Fontenele JE, de Oliveira AC, Bastos VH, Teixeira S, Rabêlo RL. Boamente: a natural language processing-based digital phenotyping tool for smart monitoring of suicidal ideation. Healthcare. 2022;10(4).

83. Mayer RE. Multimedia learning. Psychology of learning and motivation. Elsevier; 2002. p. 85–139. https://doi.org/10.1016/s0079-7421(02)80005-6

84. Belcher DD. On becoming facilitators of multimodal composing and digital design. Journal of Second Language Writing. 2017;38:80–5. https://doi.org/10.1016/j.jslw.2017.10.004

85. Sauer BA. 2. Multimodal Communication. John Wiley & Sons, Ltd.; 2011. p. 29–47.

86. Furuta H, Zen H, Schuurmans D, Faust A, Matsuo Y, Liang P. Improving dynamic object interactions in text-to-video generation with AI feedback. arXiv preprint 2024. https://arxiv.org/abs/2412.02617

87. Khoo LS, Lim MK, Chong CY, McNaney R. Machine learning for multimodal mental health detection: a systematic review of passive sensing approaches. Sensors. 2024;24(2).

88. Stahlschmidt SR, Ulfenborg B, Synnergren J. Multimodal deep learning for biomedical data fusion: a review. Brief Bioinform. 2022;23(2):bbab569. https://doi.org/10.1093/bib/bbab569 PMID: 35089332

89. Liang PP, Zadeh A, Morency LP. Foundations & trends in multimodal machine learning: principles, challenges, and open questions. ACM Computing Surveys. 2024;56(10).

90. Anisuzzaman DM, Malins JG, Friedman PA, Attia ZI. Fine-tuning large language models for specialized use cases. Mayo Clin Proc Digit Health. 2024;3(1):100184. https://doi.org/10.1016/j.mcpdig.2024.11.005 PMID: 40206998

91. Wang F, Zhang Z, Zhang X, Wu Z, Mo T, Lu Q. A comprehensive survey of small language models in the era of large language models: techniques, enhancements, applications, collaboration with LLMs, and trustworthiness. 2024. https://arxiv.org/abs/2411.03350

92. Huang L, Yu W, Ma W, Zhong W, Feng Z, Wang H. A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions. ACM Transactions on Information Systems. 2025;43(2).

93. Navigli R, Conia S, Ross B. Biases in large language models: origins, inventory, and discussion. J Data and Information Quality. 2023;15(2).

94. Adler DA, Stamatis CA, Meyerhoff J, Mohr DC, Wang F, Aranovich GJ, et al. Measuring algorithmic bias to analyze the reliability of AI tools that predict depression risk using smartphone sensed-behavioral data. Npj Ment Health Res. 2024;3(1):17. https://doi.org/10.1038/s44184-024-00057-y PMID: 38649446

95. Bouguettaya A, Stuart EM, Aboujaoude E. Racial bias in AI-mediated psychiatric diagnosis and treatment: a qualitative comparison of four large language models. NPJ Digit Med. 2025;8(1):332. https://doi.org/10.1038/s41746-025-01746-4 PMID: 40467886

96. Rzadeczka M, Sterna A, Stoli⊠ska J, Kaczy⊠ska P, Moskalewicz M. The efficacy of conversational AI in rectifying the theory-of-mind and autonomy biases: comparative analysis. JMIR Ment Health. 2025;12:e64396. https://doi.org/10.2196/64396 PMID: 39919295

97. Ali R, Tang OY, Connolly ID, Abdulrazeq HF, Mirza FN, Lim RK, et al. Demographic representation in 3 leading artificial intelligence text-to-image generators. JAMA Surg. 2024;159(1):87–95. https://doi.org/10.1001/jamasurg.2023.5695 PMID: 37966807

98. Iniesta R. The human role to guarantee an ethical AI in healthcare: a five-facts approach. AI Ethics. 2023;5(1):385–97. https://doi.org/10.1007/s43681-023-00353-x

99. de Oliveira AC, Azevedo JPC, Ruback L, Moreira R, Teixeira SS, Teles AS. Effect of explainable artificial intelligence on trust of mental health professionals in an AI-based system for suicide prevention. IEEE Access. 2025;13:60987–1005. https://doi.org/10.1109/access.2025.3556245

100. Gardiner H, Mutebi N, Parliamentary Office of Science and Technology. AI and mental healthcare - ethical and regulatory considerations. 2025.

101. Blease C, Rodman A. Generative artificial intelligence in mental healthcare: an ethical evaluation. Current Treatment Options in Psychiatry. 2024;12(1):5.

102. Zheng Y, Gan W, Chen Z, Qi Z, Liang Q, Yu PS. Large language models for medicine: a survey. Int J Mach Learn & Cyber. 2024;16(2):1015–40. https://doi.org/10.1007/s13042-024-02318-w

103. Wang W, Chen W, Luo Y, Long Y, Lin Z, Zhang L, et al. Model compression and efficient inference for large language models: a survey. arXiv preprint 2024. https://arxiv.org/abs/2402.09748

104. Stade EC, Stirman SW, Ungar LH, Boland CL, Schwartz HA, Yaden DB, et al. Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation. Npj Ment Health Res. 2024;3(1):12. https://doi.org/10.1038/s44184-024-00056-z PMID: 38609507