U.PORTO

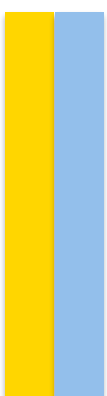
MESTRADO INTEGRADO EM MEDICINA
PSIQUIATRIA

Generative Artificial Intelligence in Mental Health

Filipa da Cruz Figueiredo

M

2026



Generative Artificial Intelligence in Mental Health

Revisão Bibliográfica

Dissertação de candidatura ao grau de Mestre em Medicina

Submetida ao Instituto de Ciências Biomédicas Abel Salazar
da Universidade do Porto

5 de junho de 2026

Estudante: Filipa da Cruz Figueiredo
Instituto de Ciências Biomédicas Abel Salazar, Universidade do Porto
Endereço eletrónico: up202005025@up.pt

Orientadora: Professora Doutora Liliana Correia de Castro
Assistente Hospitalar Graduada de Psiquiatria, Departamento de Saúde Mental, Hospital Magalhães Lemos, Unidade Local de Saúde Santo António
Afiliação: Professora Auxiliar Convidada de Psiquiatria do Mestrado integrado em Medicina, Instituto Ciências Biomédicas Abel Salazar, Universidade do Porto
Endereço eletrónico: u61159@chporto.min-saude.pt

Coorientador: Dr. Filipe Freitas Pinto
Médico Psiquiatra
Diretor Médico da Knok Healthcare
Afiliação: Professor Assistente Convidado, Faculdade de Medicina da Universidade do Porto
Endereço eletrónico: filipe.fr.pinto@gmail.com

Declaração de Integridade

Porto, 5 de junho de 2026

Assinado por: Filipa da Cruz Figueiredo
Identificação: BI14459902
Data: 2026-06-06 às 14:39:42

Filipa da Cruz Figueiredo

Mestranda e Autora

Assinado por: **LILIANA CORREIA DE CASTRO**
Num. de Identificação: 11778923
Data: 2026.06.06 14:24:38 +0100

CARTÃO DE CIDADÃO

Prof. Doutora Liliana Correia de Castro

Médica e Orientadora

Assinado por: Filipe Jorge Freitas Pinto
Num. de Identificação: BI13194959
Data: 06-06-2026 17:22:25 +01:00

CHAVE MÓVEL

Dr. Filipe Freitas Pinto

Médico e Coorientador

Acknowledgements

Completing this dissertation represents a significant milestone in my academic and personal journey, marking the end of a demanding yet meaningful stage of my medical training. It has allowed me to develop greater scientific rigour, critical thinking, and an increased appreciation for the value of knowledge, responsibility, and human relationships in Medicine.

I would like to convey my sincere gratitude to my supervisor, Professor Liliana Castro, PhD, for her availability and scientific support during the development of this dissertation. Her advice, constructive feedback, and encouragement were vital to completing this work.

I am equally grateful to my co-supervisor, Dr Filipe Pinto, for his consistent support, availability, and valuable assistance throughout this process. His considerate feedback, attention to detail, and contributions to the development of this dissertation were greatly appreciated.

I would also like to thank my family for their constant support, encouragement, and understanding throughout my academic path.

I am grateful to my friends and colleagues for their companionship and motivation, which have made this journey more pleasant and meaningful.

Finally, I would like to thank all the professors and clinicians from whom I have had the privilege of learning during my medical training. Their devotion, expertise, and example have greatly influenced my education and inspired the physician I aspire to become.

Resumo

As perturbações mentais constituem atualmente um dos principais desafios em saúde pública, devido ao seu impacto significativo na incapacidade, na qualidade de vida e na carga global de doença. Paralelamente, os avanços recentes em inteligência artificial generativa (generative artificial intelligence - GenAI), especialmente através dos large language models (LLMs), têm impulsionado o desenvolvimento de ferramentas digitais inovadoras com potencial aplicação em contextos de saúde mental. Estas tecnologias possibilitam a geração de respostas em linguagem natural, a interpretação de informações complexas e o estabelecimento de interações contextualizadas, abrindo novas perspetivas para avaliação, monitorização e intervenção psiquiátrica e psicológica.

O objetivo principal desta revisão narrativa é sistematizar a evidência científica disponível sobre as aplicações da GenAI em saúde mental, analisando o seu potencial clínico, limitações, desafios éticos e perspetivas futuras. Para isso, foi realizada uma pesquisa bibliográfica nas bases de dados PubMed, Scopus e Web of Science, incluindo estudos publicados entre janeiro de 2020 e setembro de 2025. Foram utilizados termos como ”Generative Artificial Intelligence”, ”Large Language Models”, ”Chatbots”, ”Mental Health”e ”Psychiatry”. Foram incluídos estudos empíricos que abordassem aplicações de GenAI em contextos de saúde mental envolvendo doentes, profissionais de saúde ou a população geral.

Os estudos analisados apresentam resultados promissores em diversos domínios, incluindo avaliação e triagem de sintomas, apoio à decisão clínica, psicoeducação, documentação médica automatizada e intervenções terapêuticas baseadas em chatbots. Em particular, chatbots fundamentados em terapia cognitivo-comportamental demonstraram benefícios na redução de sintomas depressivos, ansiedade, solidão e sofrimento psicológico em diferentes populações. Além disso, alguns LLMs evidenciaram capacidades relevantes em tarefas de raciocínio clínico, formulação psicodinâmica, reconhecimento emocional e geração de respostas empáticas.

No entanto, a evidência disponível permanece limitada devido à heterogeneidade metodológica, à dimensão reduzida das amostras e à predominância de estudos exploratórios conduzidos em ambientes simulados ou altamente controlados. Persistem preocupações quanto a alucinações, vieses algorítmicos, privacidade e segurança de dados, transparência dos modelos e risco de geração de conteúdos inadequados ou potencialmente prejudiciais, especialmente em contextos de maior vulnerabilidade psicológica. Acresce que a maioria das ferramentas atualmente disponíveis carece de validação clínica robusta e certificação regulatória adequada.

Apesar dessas limitações, a GenAI apresenta potencial para ampliar a acessibilidade, personalização e eficiência dos cuidados em saúde mental. A sua integração clínica deverá, contudo, ocorrer de forma cautelosa, supervisionada e fundamentada em evidência científica robusta. São essenciais futuros estudos multicêntricos, longitudinais e em contexto real, além do desenvolvimento de estruturas éticas e regulamentares apropriadas.

Palavras-chave: Inteligência Artificial Generativa; Large Language Models; Chatbots; Agentes Conversacionais; Saúde Mental; Psiquiatria; Doença Mental; Perturbações Mentais; Psicoterapia.

Abstract

Mental disorders currently constitute one of the major public health challenges, due to their significant impact on disability, quality of life, and the global burden of disease. In parallel, recent advances in generative artificial intelligence (GenAI), particularly through large language models (LLMs), have driven the development of innovative digital tools with potential applications in mental health settings. These technologies enable the generation of natural language responses, the interpretation of complex information, and the establishment of contextualised interactions, opening new perspectives for psychiatric and psychological assessment, monitoring, and intervention.

The primary objective of this narrative review is to systematise the available scientific evidence regarding the applications of GenAI in mental health, analysing its clinical potential, limitations, ethical challenges, and future perspectives. For this purpose, a literature search was conducted in the PubMed, Scopus, and Web of Science databases, including studies published between January 2020 and September 2025. Search terms included ”Generative Artificial Intelligence”, ”Large Language Models”, ”Chatbots”, ”Mental Health”, and ”Psychiatry”. Empirical studies addressing GenAI applications in mental health settings involving patients, healthcare professionals, or the general population were included.

The analysed studies demonstrated promising results across several domains, including symptom assessment and triage, clinical decision support, psychoeducation, automated medical documentation, and chatbot-based therapeutic interventions. In particular, chatbots grounded in cognitive behavioural therapy demonstrated benefits in reducing depressive symptoms, anxiety, loneliness, and psychological distress across different populations. Furthermore, some LLMs demonstrated relevant capabilities in tasks involving clinical reasoning, psychodynamic formulation, emotional recognition, and the generation of empathic responses.

However, the available evidence remains limited due to methodological heterogeneity, small sample sizes, and the predominance of exploratory studies conducted in simulated or highly controlled environments. Concerns persist regarding hallucinations, algorithmic bias, data privacy and security, model transparency, and the risk of generating inappropriate or potentially harmful content, particularly in contexts of increased psychological vulnerability. Moreover, most currently available tools lack robust clinical validation and appropriate regulatory certification.

Despite these limitations, GenAI presents considerable potential to improve the accessibility,

personalisation, and efficiency of mental healthcare. Nevertheless, its clinical integration should occur cautiously, under appropriate supervision, and grounded in robust scientific evidence. Future multicentre, longitudinal, and real-world studies are essential, alongside the development of appropriate ethical and regulatory frameworks.

Keywords: Generative Artificial Intelligence; Large Language Model; Chatbot; Conversational Agent; Mental Health; Psychiatry; Mental Illness; Mental Disorder; Psychotherapy.

List of Abbreviations

AI	Artificial Intelligence
ACC/AHA	American College of Cardiology/American Heart Association
AFP	Alpha-fetoprotein
ASD	Autism Spectrum Disorder
AUC	Area Under the Curve
BERT	Bidirectional Encoder Representations from Transformers
CAs	Conversational Agents
CBT	Cognitive Behavioural Therapy
CT	Computed Tomography
DNN	Deep Neural Network
EHR	Electronic Health Record
FT	Fine-Tuning
GAD-7	Generalized Anxiety Disorder-7
GenAI	Generative Artificial Intelligence
GPT	Generative Pre-trained Transformer
GRADE	Grading of Recommendations Assessment, Development and Evaluation
HCC	Hepatocellular Carcinoma
LEAS	Levels of Emotional Awareness Scale
LLM	Large Language Model
LSTM	Long Short-Term Memory
LVEF	Left Ventricular Ejection Fraction
MDR	Medical Device Regulation
MeSH	Medical Subject Headings
ML	Machine Learning
MRI	Magnetic Resonance Imaging
NER	Named Entity Recognition
NLP	Natural Language Processing
PEMAT-P	Patient Education Materials Assessment Tool for Printable Materials
PHQ-9	Patient Health Questionnaire-9
RAG	Retrieval Augmented Generation
RAS	Robot-Assisted Surgery
SEER	Surveillance, Epidemiology, and End Results
SLAM	Simultaneous Localization and Mapping
TNM	Tumour-Node-Metastasis
UCLA	University of California, Los Angeles
UMLS	Unified Medical Language System
VB-MAPP	Verbal Behavior Milestones Assessment and Placement Program
VGG	Visual Geometry Group
XGBoost	Extreme Gradient Boosting

Contents

List of Tables

List of Figures

1. Introduction

Artificial intelligence (AI) has become a transformative force across healthcare, with well-established applications in areas such as diagnostics, drug discovery, and clinical decision-making. However, given the significant global burden of psychiatric conditions, it is important to consider how far these advances extend to mental healthcare [1].

Compared with other medical fields, the application of AI in mental health is characterised by considerable heterogeneity, reflecting the wide range of approaches and varying levels of clinical integration. This raises the question of its potential role in addressing existing gaps in psychiatric care and whether similar benefits can be achieved in this domain [2].

Currently, healthcare is undergoing a paradigm shift driven by the emergence of generative artificial intelligence (GenAI), a class of models capable not only of analysing data but also of generating new content across multiple modalities, including text, images or voice modalities [3]. Unlike traditional AI approaches, which are primarily focused on prediction or classification tasks, GenAI systems are designed to produce novel, contextually relevant outputs, thereby expanding their applicability across a wide range of clinical and non-clinical settings.

Among the most widely adopted forms of GenAI in healthcare are large language models (LLMs), such as Generative Pre-trained Transformer (GPT)-based systems, which offer unique opportunities for communication, interaction, and knowledge synthesis [3,4]. These models are typically built upon transformer architectures and are trained on vast amounts of textual data, enabling them to understand and generate human-like language. In contrast to earlier natural language processing (NLP) techniques, LLMs allow for dynamic, prompt-based interaction, facilitating more natural and context-aware exchanges between humans and machines [5,6].

LLMs represent the culmination of decades of progress in machine learning (ML) and NLP. ML, now considered a cornerstone of modern AI, enables systems to learn patterns from both structured and unstructured data [7]. In healthcare, conventional ML approaches, such as supervised, unsupervised, and reinforcement learning have been extensively applied in areas including diagnostic imaging, prognostic modelling, and risk prediction [8]. However, these methods present important limitations, particularly their dependence on labelled datasets and their limited capacity to process complex, unstructured human communication. This limitation is especially relevant in fields such as mental health, where clinical information is often derived from subjective narratives,

speech, and written text rather than objective biomarkers[7,8].

In contrast, LLMs introduce more flexible learning paradigms, including transfer learning, as well as few-shot and zero-shot learning, allowing them to generalise across tasks with limited or no labelled data[5]. This capability enables the adaptation of a single model to multiple clinical tasks, improving scalability and efficiency. Nevertheless, general-purpose LLMs, such as ChatGPT, are not specifically designed for clinical use, raising important concerns related to hallucinations, bias, and patient privacy. These limitations highlight the need for domain-specific models that are explicitly tailored to the complexity and requirements of healthcare environments[5]. In this context, LLMs can be customised to clinical tasks relying on two main methodological strategies: fine-tuning (FT) and retrieval augmented generation (RAG).

Instead of training a model from the very start, FT offers a low-resource approach by optimizing a pre-trained model through additional training on a limited, task-oriented dataset[9]. As an example, Clinical (Bio)BERT (Bidirectional Encoder Representations from Transformers) relies on this process, as it is fine-tuned on specialised medical corpora which is crucial to perform biomedical NLP tasks. It arises from the BERT architecture which is described as a Deep Learning (DL) architecture that led to the emergence of modern LLMs[9]. Apart from Clinical (Bio) BERT, other models addressing the biomedical domain were developed, including BioBERT and UMLS (Unified Medical Language System)-BERT. BioBERT is a specialised pre-trained model designed to enhance performance in biomedical text mining tasks (the process of extracting meaningful data). However, comparative analysis demonstrated that Clinical (Bio) BERT outperformed BERT and BioBERT in 3 of the 5 evaluated NLP tasks. Regarding UMLS-BERT, this model was designed to integrate formalised expert knowledge from the UMLS at the time it is pre-trained. This process creates clinically informative input representations, enabling it to achieve superior performance compared to other domain-specific models across tasks involving named entity recognition (NER) and clinical NLP inference[10]. More recently, MedLM has emerged as a large-scale clinical language model, extending these approaches through the use of larger datasets and further adaptation techniques, including FT, to support tasks such as medical question answering and clinical summarisation[11] (Table I).

RAG is used to approach an important challenge of LLMs, which is related to the possibly outdated training data of general models[9]. Considering this, RAGs' technology strengthens LLMs by accessing external sources of knowledge[9,12]. This approach enables the model to ground its responses in the most accurate, up-to-date, and evidence-based information, while also incorporating contextual knowledge from external sources, such as institutional guidelines or local databases[9].

This fact is particularly relevant for real-time decision-making systems[12], and it has been utilised to synthesise and retrieve main clinical information from electronic health records (EHRs)[5,13]. Its capacity of improving accuracy in evidence-based medicine has been investigated and compared to non-customised models (Figure 1).

These approaches significantly improve the accuracy and thus, the clinical applicability, providing context-aware and more reliable outputs. Overall, these models are tailored to the specific language and complexity of medical data, connecting technological innovation with clinical applicability and contributing to more effective and valuable healthcare outcomes[2].

1.1. AI applications in Medical Fields

Over the past decade, AI has rapidly evolved from early ML models focused on structured data into a transformative tool widely applied across healthcare for clinical decision support, diagnosis, and patient monitoring[8,14]. Given the strong evidence for AI in imaging, particularly in radiology, as well as its wide-ranging applications across other medical specialties, three other key medical broad domains were selected to illustrate its clinical impact: cardiovascular medicine, oncology, and surgery.

Notable advances have been made in cardiovascular medicine, where AI is extensively applied to advance diagnostic accuracy, risk stratification, and outcome prediction. Regarding diagnosis, AI algorithms assist clinicians by generating automated quantitative cardiac evaluations[15]. DL approaches, especially Convolutional Neural Networks (CNNs) (a class of models specialised in image analysis), can automatically analyse echocardiograms to outline the left ventricle and precisely estimate the left ventricular ejection fraction (LVEF), achieving high performance (Dice coefficient ≈ 0.92, a measure of overlap between predicted and reference segmentations). Deep neural networks (DNNs), which are multilayered AI models capable of learning complex data patterns, have been shown to accurately distinguish multiple arrhythmia types, reaching an area under the curve (AUC) of 0.97, comparable to expert cardiologists. CNN architectures are also used to detect left-ventricular dysfunction and valvular abnormalities using electrocardiogram and phonocardiogram data (heart sounds recorded by digital stethoscopes), with accuracy rates ranging from 70% to 98%. In addition to diagnostic use, ML algorithms that combine imaging, clinical, and genomic information improve cardiovascular risk assessment and outperform conventional predictors like the Framingham or American College of Cardiology/American Heart Association scores (ACC/AHA), demonstrating performance levels approaching 90% in forecasting one-year mortality among heart-failure

patients.

In the oncology field, most notably in colon cancer studies, AI methods are mainly employed for image classification, segmentation, and outcome prediction. CNN models such as ResNet, Visual Geometry Group (VGG), and EfficientNet can differentiate benign from malignant specimens in histopathological images achieving accuracies of up to 99.9%. U-Net architectures, a DL framework optimised for pixel-level image segmentation, are employed to outline tumour borders, glands and nuclei, enabling automated recognition of malignant regions. Concerning prognostic, ML algorithms that include Random Forest and eXtreme Gradient Boosting (XGBoost) analyse extensive data sources, such as those from the Surveillance, Epidemiology, and End Results (SEER) program to estimate survival and recurrence probabilities across different time spans[16]. In addition, in hepatocellular carcinoma (HCC), hybrid AI approaches that integrate molecular biomarkers (e.g., long non-coding RNA,lncRNA, signatures) with clinical variables like alpha-fetoprotein (AFP) levels and tumour-node-metastasis (TNM) staging, enhance the prediction of early recurrence and treatment response, contributing to precision oncology[17].

Within the surgical domain, AI contributes across all stages of care, enhancing preoperative decision-making to intraoperative assistance and postoperative evaluation. Regarding preoperative imaging analysis, AI plays a crucial role, where DL models process data from radiography, computed tomography (CT), magnetic resonance imaging (MRI), and ultrasound to identify, classify, and segment anatomical structures and lesions in organs such as the lung, bladder, and breast[18]. During robot-assisted surgery (RAS), AI interprets motion and video inputs to access operator skill, identify gestures, and determine procedural phases, attaining accuracy rates above 90%[19]. Extending beyond robotic systems, AI also contributes to intraoperative guidance by utilizing CNN models capable of real-time instrument localization (precision $\approx$ 76-90%), whereas Visual Odometry (estimates camera pose for endoscopic capsule robots) and simultaneous localization and mapping (SLAM) support augmented-reality navigation and tissue motion tracking in minimally invasive surgery[18]. Postoperatively, ML models can be used to estimate outcomes, including expected hospital stay and urinary continence recovery after robotic prostatectomy, with performance levels close to 85-89%[19].

Overall, these developments highlight the broad clinical impact of AI across multiple specialties and its potential to be extended to other domains. In this context, psychiatry emerges as a particularly promising, albeit challenging, field for its application.

1.2. The Emerging Capabilities of Generative AI in Mental Health

The application of GenAI is specifically promising in mental health. This arises from the fact that psychiatric practice counts significantly on subjective experience, language, nuanced communication between patients and clinicians and scarce human resources [9,20]. Compared with imaging-based specialties, where AI is already substantially developed, the use of GenAI in mental health is still at an early but fast-evolving stage [21,22].

The constantly advancing abilities of LLMs powered by NLP holds great promise for mental healthcare, as these systems effectively recognise and respond to emotional cues [2,14]. Through techniques like sentiment analysis, NLP allows AI algorithms to examine extensive textual data from social media or other platforms to spot early signs of emotional imbalance and potential mental health issues [2,9,14]. Evidence suggests that domain-specific models perform with remarkable efficacy, achieving accuracies of up to 93.14% in sentiment analysis and 85.46% in emotion classification on benchmark datasets [9]. While sentiment analysis classifies the text into polarity groups, as positive, neutral, negative and eventually mixed, emotion analysis defines classes like ”joy”, ”sadness”, ”anger” and ”fear”.

Building on these capabilities, more advanced approaches, including attention-based long short-term memory (LSTM) models, have further improved the accuracy of emotion recognition, a critical component for delivering meaningful mental health support [23]. Emerging evidence also suggests that LLMs can demonstrate strong emotional understanding and empathetic responsiveness in controlled evaluations, reflecting their potential to simulate aspects of human-like emotional interaction [4,24,25]. This distinctive combination of emotional intelligence with 24/7 availability [9,20,21,26] highlights the transformative potential of GenAI in mental healthcare, offering continuous, non-judgmental support and playing a crucial role in improving access and reducing stigma [2,9,14,23].

These capabilities do not represent specific clinical applications per se, but rather the foundational mechanisms that enable the diverse range of solutions explored in this research work.

2. Aim of the review

This narrative review aims to investigate the current landscape of GenAI solutions applied to mental health, as well as to evaluate the effectiveness of the tools currently available, through a critical analysis of the published evidence. To provide a structured framework for analysis, AI applications were categorised according to their intended clinical function, in line with existing healthcare classification approaches that underpin regulatory frameworks such as the Medical Device Regulation (MDR) and the forthcoming AI Act [1,27]. These frameworks adopt a risk-based perspective, whereby the level of regulatory oversight is determined by the potential impact on patient safety.

Accordingly, four principal domains of application were considered: Symptom Assessment and Triage, Administrative Support, Clinical Decision Support, and Patient Support. This functional classification facilitates a more structured comparison of existing evidence across different types of AI applications [1,27]. Studies that focused mainly on Evaluating/Benchmarking AI Systems, as opposed to testing a specific clinical application, were considered separately as a complementary category.

Within this context, the present review aims not only to assess the clinical relevance and effectiveness of GenAI-based tools across these domains, but also to evaluate their potential for integration into routine clinical practice. Particular attention is given to their operational advantages, inherent limitations, and the broader implications for future development in mental healthcare. This classification was subsequently used as the organisational framework for the present review.

3. Methods

This review covers published in peer-reviewed journals from January 2020 to September 2025 addressing the application of GenAI in mental health, focusing on four domains: Symptom Assessment and Triage, Administrative Support, Clinical Decision Support, and Patient Support. Additionally, studies that assessed the performance of AI systems were considered separately as Evaluation/Benchmarking of AI Systems. The literature search was conducted between September and October 2025. Electronic databases including PubMed/MEDLINE, Scopus, and ScienceDirect were systematically searched. To account for the rapidly evolving nature of the field, the search was supplemented by screening preprints available on arXiv.

The search strategy combined Medical Subject Headings (MeSH) and keywords using Boolean operators, as follows: (”Generative Artificial Intelligence” OR ”Large Language Model” OR ”Chatbot” OR ”Conversational Agent”) AND (”Mental Health” OR ”Psychiatry” OR ”Mental Illness” OR ”Mental Disorder” OR ”Psychotherapy”).

Studies were included if they met all the following inclusion criteria:

(1) published in English;

(2) published between January 2020 and September 2025;

(3) addressed the application of generative artificial intelligence (GenAI), including large language models and chatbots, in mental health contexts;

(4) evaluated the use of GenAI in simulated, experimental, or real-world settings involving patients, healthcare professionals, or general users;

(5) reported outcomes related to at least one of the following domains: symptom assessment and triage, administrative support, clinical decision support, patient support or evaluating/benchmarking AI systems;

(6) included empirical studies, such as randomised controlled trials, observational studies, simulation-based evaluations, or qualitative studies.

Studies were excluded if they met any of the following exclusion criteria:

(1) focused exclusively on non-generative, discriminative artificial intelligence models;

(2) did not specifically relate to mental health applications;

(3) consisted solely of technical descriptions or conceptual analyses without evaluation of outcomes relevant to clinical or user-level application.

Study selection and screening were conducted by a single reviewer, who applied the predefined inclusion and exclusion criteria consistently across all stages.

4. Results

The initial search resulted in 1316 studies (Pubmed: 152, Scopus: 601, Science Direct: 563). After excluding the duplicates, 951 studies remained. Subsequently, an initial screening was conducted based only on the title, resulting in the inclusion of 403 studies. Afterwards, the abstracts of the remain articles were analysed, and 95 studies were selected for full-text assessment. In addition, cross-referencing was used to identify relevant studies that weren't documented in the baseline search. Ultimately, 20 studies were considered, as they met the inclusion criteria and the aim of the review. Priority was given to recent studies and those that demonstrated clinical applicability (Figure 2).

The included studies were methodologically heterogeneous, comprising randomised controlled trials (n = 4), simulation-based or benchmark studies (n = 5), observational or cross sectional designs (n = 5), qualitative or mixed-methods studies (n = 3), and exploratory or pilot investigations (n = 3).

Most studies focused on patient support (n = 10), with relatively few addressing symptom assessment and triage (n = 2), clinical decision support (n = 1), or administrative support (n = 1). Six studies evaluated AI system performance and benchmarking in mental health (Table II and Table III; Figure 3 and Figure 4).

4.1. Patient Support

In a randomised controlled trial conducted in Hong Kong, Chen et al. (2025) compared an AI-based chatbot with a nurse-led telephone hotline in adults seeking support for symptoms of anxiety and depression. Participants accessed the service voluntarily and were then randomly assigned to either the chatbot or the telephone intervention, both delivered via telehealth. The effectiveness of each approach was evaluated using changes in symptom severity, measured with the Patient Health Questionnaire-9 (PHQ-9) and the Generalized Anxiety Disorder-7 (GAD-7). The chatbot group showed significant reductions in both depression and anxiety scores, whereas no significant changes were observed in the hotline group[26].

Wang et al. (2025) evaluated a cognitive behavioural therapy (CBT)-based AI chatbot in a randomised controlled trial conducted in China among university students presenting with depressive symptoms and loneliness. Participants were recruited in an academic setting and randomly

assigned either to the chatbot intervention or to a waitlist control group. The intervention was delivered over a short period through an AI-driven conversational system designed to provide automated CBT-based support adapted to the local cultural context[28]. Outcomes were assessed using validated self-report instruments, including the PHQ-9 for depressive symptoms, the GAD-7 for anxiety, and a standardised loneliness scale. At the end of the intervention, participants in the chatbot group showed a significant reduction in depressive symptom scores compared with the control group ($p < 0.001$), along with improvements in loneliness. No significant differences were observed for anxiety outcomes. Greater improvements were reported among individuals experiencing higher levels of financial stress[28].

Alanezi (2024) conducted a qualitative study in Saudi Arabia to evaluate the use of ChatGPT for mental health support among adult outpatients with conditions such as anxiety and depression. Participants used ChatGPT over a two-week period, and outcomes were assessed through semi-structured interviews analysed using thematic analysis, focusing on domains such as psychoeducation, emotional support, self-management, and perceived reliability[29]. The results showed that participants frequently reported benefits in psychoeducation and symptom management, with over 60% indicating that ChatGPT improved their understanding of mental health conditions. Emotional support was also commonly described. However, concerns regarding the accuracy and reliability of the information were prominent, being reported by approximately 80% of participants, alongside consistent limitations in diagnostic and assessment capabilities[29].

Hristidis et al. (2023) compared ChatGPT with Google by submitting 60 dementia-related queries, including both informational and service-oriented questions derived from validated knowledge scales and common caregiver needs. Responses were evaluated across domains such as relevance, reliability, objectivity, currency, and readability by three domain experts using predefined categorical rating scales, with interrater agreement calculated to ensure consistency. ChatGPT responses were markedly more relevant to the queries, with 96.7% rated as highly relevant compared to 60% of Google responses. In contrast, Google performed better in terms of currency and source reliability, while ChatGPT outputs were more consistently objective[30].

Amin et al. (2023) assessed ChatGPT's (GPT-3.5) ability to provide support for vaping cessation using 10 prompts based on user posts from a Reddit quitting community. Responses were evaluated by five experts in tobacco and substance use research across domains including relevance, accuracy, clarity, quality, and empathy, each rated on a 3-point scale (1 = poor; 2 = satisfactory; 3 = excellent). Overall, responses were consistently rated close to the highest category, with

a mean score of 2.7 out of 3 for purpose, quality, and clarity, indicating that most answers were judged to be appropriate and well aligned with users' needs[31].

Sabour et al. (2023) investigated the effectiveness of Emohaa, a conversational agent (CA) designed to provide both cognitive support (based on CBT exercises) and emotional support through interactive dialogue. The study was conducted in China and included 247 participants with symptoms of mental distress, who were randomly assigned to one of three groups: two intervention groups (CBT-based Emohaa and a full version combining cognitive and emotional support) and a control group that did not receive any intervention. Outcomes, including depression, negative affect, and insomnia, were assessed using validated self-report scales, such as the Centre for Epidemiologic Studies Depression Scale and the University of California, Los Angeles (UCLA) Loneliness Scale, and analysed by comparing symptom changes across groups over the study period. Overall, participants in the intervention conditions showed significantly greater improvements in depression and negative affect compared with the control group, with statistically significant differences between groups ($p=0.002$ and $p=0.003$, respectively)[32].

Tang et al. (2024) evaluated the effectiveness of EmoEden, a Gen AI-based tool designed to support emotional learning in children with high-functioning autism, focusing on emotional recognition and expression outcomes. The intervention was assessed through a 22-day user study involving six participants, using pre- and post-intervention comparisons alongside observational measures of task performance and user engagement, rather than standardised clinical scales. Results showed improvements in emotional recognition and expression abilities across all participants following the intervention, with gains maintained for at least five days after completion[33].

Koegel et al. (2025) conducted a randomised clinical trial evaluating an AI-based conversational training programme (Noora) designed to improve empathetic verbal responses in autistic adolescents and adults (n = 29). The main parameter assessed was the proportion of correct empathetic responses during social conversation. Outcomes were evaluated through pre-post analysis of conversation samples (percentage of correct responses) and self-report questionnaires. The control group was a waitlist condition without intervention. Within the intervention, 79% of participants improved or maintained accuracy when comparing their first and last 50 responses[34].

Deng et al. (2024) evaluated ASD-Chat, a dialogue-based intervention system for children diagnosed with autism spectrum disorder (ASD) grounded in the Verbal Behavior Milestones Assessment and Placement Program (VB-MAPP) and powered by an LLM. The system was tested in a sample of children with autism undergoing social communication training, targeting language and

interaction skills through structured conversational exchanges[35]. Outcomes were assessed using multimodal measures, including behavioural indicators (engagement and response quality derived from text and speech) and physiological data (functional near-infrared spectroscopy). Compared with professional interventionists, children interacting with ASD-Chat demonstrated higher engagement, while response quality was slightly lower. Importantly, physiological responses were highly similar across conditions, with differences in brain activation of approximately 0.1, indicating comparable intervention effects between ASD-Chat and human-delivered interventions[35].

McFayden et al. (2024) evaluated the quality of ChatGPT-4 responses to 13 common caregiver questions about autism, covering three domains: basic information, myths/misconceptions, and resources. The model's outputs were assessed across multiple parameters, including correctness, clarity, and conciseness (the "3Cs"), alongside language use, understandability and actionability (using the validated Patient Education Materials Assessment Tool for Printable Materials, PEMAT-P, tool), and reference accuracy. Responses were rated on structured scales (e.g., 1-4 for the 3Cs and percentage scores for PEMAT-P domains) by independent evaluators with inter-rater reliability analysis. Overall, ChatGPT demonstrated high-quality informational performance, with mean 3C scores of approximately 3.6-3.8 out of 4 across domains, indicating responses were largely accurate, clear, and concise. However, reference accuracy was low, with fewer than half of hyperlinks being functional[36].

4.2. Tools for Symptom Assessment and Triage

D'Souza et al. (2023) evaluated ChatGPT-3.5 in a simulated clinical study using 100 psychiatric case vignettes from a standardised casebook (*100 Cases in Psychiatry*). The model was assessed across multiple parameters, including diagnosis, differential diagnosis, assessment, investigation, management, counselling, clinical reasoning, ethical reasoning, and prognosis. Responses were evaluated by two psychiatry faculty members against reference answers using a grading system (Grade A: correct and comprehensive; Grade B: mostly correct; Grade C: partially correct). Overall, the majority of responses were rated within the highest categories (Grade A or B: 92%)[3].

The performance and safety of ChatGPT in mental health assessment were examined through a simulation study using three hypothetical patient scenarios involving insomnia with increasing clinical complexity (Dergaa et al., 2024). The responses were qualitatively evaluated by two academic clinicians (a psychiatry professor and a practicing psychiatrist), considering clinical adequacy, appropriateness of management, and safety, without the use of standardised rating scales or quan-

titative metrics. In the least complex case, the outputs were generally appropriate but lacked specificity. In contrast, in more complex cases, the outputs were judged to be clinically inadequate, particularly due to the absence of follow-up questions to clarify key aspects of the patient's condition and the provision of generalised recommendations that did not account for clinical severity or potential comorbidities. No numerical measures of accuracy were reported[6].

4.3. Administrative Support

Blease et al. (2024) investigated psychiatrists' experiences and views on the use of LLM-based chatbots through an online survey of 138 respondents. The study examined patterns of use, perceived impact on clinical practice, and expectations regarding both professional and patient use of these tools. Responses were collected through structured survey questions alongside free-text comments. A key finding was the presence of mixed attitudes, with clinicians recognising administrative benefits while raising concerns about reliability, bias, and patient safety. The most consistent finding was related to documentation, with approximately 70% of respondents agreeing that these tools could improve efficiency in clinical record-keeping[37].

4.4. Evaluation / Benchmarking of AI Systems

The performance of LLMs in simulated clinical communication was evaluated using a benchmarking framework based on multi-turn mental health interactions (Pombal et al., 2025). The study did not include real patients but generated conversations between a clinician model and a simulated patient model, based on predefined profiles incorporating demographic and psychosocial characteristics. Interactions were assessed using a structured framework covering five domains: clinical accuracy and competence, ethical and professional conduct, assessment and response, therapeutic relationship, and AI-specific communication quality[38]. Across 12 large LLMs, mean performance scores remained below 4 on a 6-point scale (where 6 represented the highest level of performance) across all domains. Lower scores were observed in scenarios involving prolonged interactions and more severe depressive or anxiety symptoms. Overall, scores remained within the mid-range of the scale across evaluation criteria[38].

Herrmann-Werner et al. (2024) examined the performance of GPT-4 using 307 multiple-choice questions in psychosomatic medicine, a field integrating psychological and physical aspects of illness, with analysis across cognitive levels defined by Bloom's taxonomy. Responses were

quantitatively evaluated based on accuracy and qualitatively analysed to classify reasoning errors according to cognitive domains. Overall, GPT-4 achieved a high level of performance, correctly answering approximately 93% of questions. However, errors were predominantly observed at lower cognitive levels, particularly in factual recall and basic understanding, rather than in more complex reasoning tasks[39].

Hadar-Shoval et al. (2023) examined ChatGPT's capacity for mentalizing, that is, its ability to represent and interpret mental states, by assessing whether its responses could be adapted to different personality structures, specifically Borderline Personality Disorder and Schizoid Personality Disorder. The model's outputs were evaluated in terms of emotional awareness using the Levels of Emotional Awareness Scale (LEAS), a validated framework that measures the complexity and differentiation of emotional understanding. ChatGPT demonstrated the ability to generate responses that appropriately reflected the distinct emotional profiles of each condition, producing more intense and nuanced descriptions for borderline presentations and more detached ones for schizoid traits, with LEAS scores reaching the maximum possible value of 80 versus 47, respectively[40].

He et al. (2024) evaluated the performance of ChatGPT-4 in responding to web-based questions from autistic patients, comparing its answers with those provided by physicians and another language model. The analysis was based on 100 patient cases comprising a total of 239 autism-related questions. The responses were assessed across multiple dimensions, including relevance, accuracy, usefulness, and empathy. Three senior physicians rated each response using a structured 5-point Likert scale and also indicated their preferred answer among the options. Overall, ChatGPT demonstrated performance comparable to physicians in terms of relevance and accuracy, and notably achieved higher ratings in empathy. However, physicians' responses were more frequently preferred overall (46.9% vs 34.9% for ChatGPT), suggesting that while ChatGPT provides clinically valuable and empathetic answers, it does not yet consistently surpass human experts in overall quality[41].

Sezgin et al. (2023) assessed the clinical accuracy of LLM responses to 14 patient-focused questions on postpartum depression, comparing ChatGPT (GPT-4), Bard and Google Search. Responses were rated by two physicians against guideline-based answers from the American College of Obstetricians and Gynaecologists using a scale based on the Grading of Recommendations Assessment, Development and Evaluation (GRADE) framework, ranging from 0 (no or incorrect response) to 4 (fully accurate and comprehensive), with high interrater agreement. In this framework, ChatGPT achieved the highest overall performance, with a mean score of 3.93 out of 4, sig-

nificantly outperforming both Bard and Google Search[42].

Hwang et al. (2024) conducted an exploratory study to assess ChatGPT's ability to generate psychodynamic formulations based on a clinical case previously published in a psychoanalytic journal. The model was prompted using inputs with varying levels of background information, and the resulting formulations were evaluated by five psychiatrists. Each output was rated as either appropriate or not appropriate, and the different versions were also compared in terms of overall quality. The findings showed that formulations generated using psychodynamic concepts, as well as those based on a simple case description, were considered appropriate by all evaluators, whereas performance declined when keyword-based prompts were used (appropriateness: 80% and 40%, respectively)[43].

4.5. Clinical Decision Support

Elyoseph et al. (2024) conducted a vignette-based study assessing the ability of four LLMs (ChatGPT-3.5, ChatGPT-4, Claude and Bard) to generate free-text prognostic responses in depression. For each case, the models produced textual predictions regarding expected outcomes with and without treatment, as well as possible long-term consequences. These outputs were compared with responses from psychiatrists, clinical psychologists, mental health nurses, general practitioners, and members of the general public, with all responses categorised into prognostic outcomes (improvement, no change or deterioration)[44]. All four LLMs consistently recognised depression as the main condition and recommended a combination of psychotherapy and antidepressant medication. Differences were observed in the distribution of prognostic categories: ChatGPT-3.5 generated a greater proportion of responses indicating deterioration, whereas ChatGPT-4, Claude and Bard produced response patterns more closely aligned with those of human participants, particularly in predicting no improvement or worsening in the absence of treatment. Exact numerical proportions were not reported, and comparisons were based on relative differences in response frequencies across categories[44].

5. Discussion

5.1. Key findings

This review identifies patient support via CAs and chatbots as the most extensively studied application of GenAI in mental healthcare[4,9]. The included studies primarily focused on systems designed to deliver psychoeducation, emotional support, CBT-based interventions, symptom monitoring, and social or motivational guidance. Evidence from randomised controlled trials and exploratory studies indicates that chatbot-based interventions may improve a broad spectrum of psychiatric symptoms, such as depression, anxiety, loneliness, insomnia, stress-related symptoms, emotional dysregulation, and social communication difficulties in ASD[26,28,32–34]. Additional reported benefits include enhanced emotional awareness, empathy training, support for vaping cessation, and psychoeducational guidance for both caregivers and patients.

Despite these promising results, the current evidence base is limited by significant methodological constraints. Most intervention studies compared chatbot-based systems with waiting-list controls, no intervention, or low-intensity comparators such as self-help materials, rather than with standard psychiatric or psychological care[26,28,32]. Therefore, while these tools may reduce symptom burden in certain populations, the evidence does not support their equivalence or superiority to mental health professionals.

Still, the findings indicate that GenAI systems may serve a valuable complementary function in addressing treatment gaps, particularly for populations experiencing barriers to conventional mental healthcare, such as financial limitations, stigma, geographical isolation, or shortages of specialised professionals. The scalability, continuous availability, and relatively low operational costs of these systems may facilitate broader access to low-intensity mental health support. However, they should not be regarded as substitutes for clinicians, especially in cases of severe or complex psychiatric disorders where diagnostic nuance, therapeutic alliance, and risk management are critical[4,6,37,45].

The observed benefits were relatively consistent across various chatbot architectures and underlying language models. The reviewed studies encompassed systems based on fine-tuned mental health models, specialised CBT-oriented CAs, and commercial LLMs such as GPT-based systems. Of note, there is no clear evidence that specialised or fine-tuned psychiatric models consistently outperform general-purpose commercial systems[38,42,44]. This suggests that factors such as

interaction design quality, therapeutic framework, and implementation strategy may currently be more influential than the specific model architecture.

Simultaneously, concerns regarding safety and unintended psychological effects are increasingly prominent. Recent evidence indicates that some users may develop emotionally dependent relationships with CAs and may follow chatbot recommendations uncritically, even when these are potentially harmful. Of particular concern are reports that adversarial prompting or so-called ”jailbreaking” techniques can bypass built-in safety guardrails, resulting in models generating detailed self-harm or suicide-related instructions[37,46].

These findings underscore the necessity for comprehensive testing, robust monitoring systems, and ongoing adversarial evaluation prior to deploying GenAI systems in sensitive mental health contexts. Although many commercial models incorporate safeguards to prevent harmful outputs, current evidence suggests these protections are imperfect and susceptible to manipulation[37]. Ethically and clinically, this raises significant concerns regarding the principle of non-maleficence (”do no harm”), especially when these technologies are accessed by psychologically vulnerable individuals.

In the area of clinical decision support and diagnostic reasoning, the evidence identified in this review is comparatively limited. Nonetheless, available studies and benchmarking approaches suggest that LLMs can perform well when responding to psychiatric case vignettes, exam-style questions, or structured diagnostic scenarios[38,39,43]. These findings indicate that GenAI systems can synthesise structured clinical information and generate coherent psychiatric interpretations under controlled conditions.

These results should be interpreted with caution, as most studies assessing diagnostic performance relied on pre-constructed clinical vignettes, where relevant symptoms, patient history, and contextual information were already organised and curated[3,29,38]. As a result, strong performance in these structured tasks may overestimate the actual clinical capabilities of GenAI systems in routine psychiatric practice. In real-world psychiatric practice, the primary challenge often lies not in diagnostic reasoning, but in psychiatric semiology: eliciting accurate information through nuanced communication, interpreting subjective narratives, identifying inconsistencies, and establishing rapport with patients. Psychiatric interviews demand flexibility, emotional attunement, and sensitivity to social and cultural context, especially as patients may provide incomplete, contradictory, or emotionally charged information. These interpersonal and contextual aspects remain difficult to replicate with current AI systems.

Evidence regarding the use of GenAI in psychiatric triage is also limited. Preliminary studies suggest that CAs may assist with symptom screening, information provision, and prioritisation of care pathways. However, current evidence is insufficient to support widespread implementation. Given the risks of underestimating psychiatric severity, particularly in cases involving suicidality, psychosis, or severe mood disorders, further research evaluating the reliability, safety, and real-world performance of AI-supported triage systems is urgently required[29].

In summary, the findings of this review indicate that GenAI technologies demonstrate significant potential in mental healthcare, particularly as scalable tools for increasing access to psychoeducation and low-intensity psychological support. However, current evidence does not justify their use as substitutes for trained mental health professionals. GenAI systems should be regarded as adjunctive technologies, with implementation that is carefully supervised, clinically validated, and ethically regulated.

5.2. Gaps in the Current Evidence

Although the literature on GenAI in mental health is expanding rapidly, several significant gaps persist within the current evidence base (Figure 5). Most studies to date have been conducted under highly controlled or simulated conditions rather than within real-world psychiatric settings. A large proportion of the reviewed evidence is based on clinical vignettes, benchmark datasets, predefined prompts, or short-term experimental interventions[3,6,38]. Although these designs facilitate preliminary evaluation, they do not capture the complexity of routine psychiatric care, which involves dynamic interpersonal interactions, incomplete information, and longitudinal follow-up. As a result, the current literature offers limited evidence regarding the clinical effectiveness, safety, and workflow integration of GenAI systems in everyday mental healthcare practice.

None of the identified studies evaluated GenAI tools that were formally certified as medical devices or implemented within regulated clinical infrastructures. Most interventions utilised experimental chatbot systems or commercially available LLMs deployed outside of formally approved healthcare environments[37,47]. This gap is especially relevant considering increasing regulatory scrutiny, such as the European AI Act and Software as a Medical Device framework. The lack of studies involving clinically certified systems restricts the applicability of current findings to routine healthcare implementation[1].

A further limitation is the restricted geographical and cultural diversity represented in the

available evidence. Most studies have been conducted in high-income or region-specific contexts, particularly in China, the United States, and selected English-speaking settings. Additionally, many interventions were developed and evaluated primarily in English-language environments. Given that psychiatric assessment and emotional expression are shaped by linguistic, social, and cultural factors, the generalisability of current findings to diverse populations remains uncertain[26,28,29]. Cross-cultural and multi-region studies are therefore a critical unmet need in this field.

Although several studies have compared the performance of different LLMs, benchmarking approaches remain highly heterogeneous. Existing evaluations have used varying clinical tasks, scoring systems, prompts, and outcome measures, which complicates direct comparison between models[38,41,42]. Furthermore, many studies have focused primarily on isolated performance metrics, such as diagnostic accuracy or empathy ratings, without evaluating broader dimensions such as safety, consistency, longitudinal reliability, or clinical utility. The absence of standardised psychiatric benchmarking frameworks currently limits the ability to determine which models are most appropriate for specific mental health applications[38].

6. Conclusion

GenAI is increasingly explored in mental healthcare, with current applications including symptom assessment and triage, clinical decision support, administrative assistance, and patient support interventions. The studies reviewed indicate that LLMs and CAs demonstrate potential utility in several areas of psychiatric care, particularly in psychoeducation, emotional support, psychotherapeutic guidance, and selected clinical reasoning tasks. Intervention studies have reported improvements in outcomes such as depression, loneliness, insomnia, emotional distress, and emotional awareness following interaction with AI-based systems[28,32,44].

Multiple studies have demonstrated that GenAI systems can produce coherent, contextually relevant, and empathetic responses in mental health-related interactions. In specific tasks such as emotional awareness or empathy evaluations, certain models achieved ratings comparable to or occasionally exceeding those of human evaluators[41,44]. Exploratory studies further suggest that these systems may assist in psychodynamic formulation[43], medical question answering[30,42], and support for neurodevelopmental conditions such as ASD[30,34,35,42,43]. However, performance varied considerably depending on the task, prompting strategy, and evaluation framework employed.

Despite these promising findings, the current evidence base is limited and heterogeneous. Many studies were conducted in simulated or experimental settings, often involving small sample sizes, short intervention periods, or highly selected populations. Real-world clinical validation is scarce, and few studies have evaluated the long-term effectiveness, safety, or implementation of these systems in routine psychiatric care. Additionally, methodological differences between studies, including variations in AI models, outcome measures, and evaluation criteria, hinder direct comparison and limit the generalisability of findings[37,39].

This review highlights several significant limitations and risks associated with the use of GenAI in mental health. Concerns regarding hallucinations, inaccurate or potentially harmful responses, algorithmic bias, privacy and confidentiality, lack of transparency, and insufficient regulatory oversight remain substantial barriers to clinical implementation[37,39]. These issues are particularly relevant in psychiatry, where interactions often involve vulnerable individuals and high-risk clinical scenarios. Recent evidence also indicates that certain safety mechanisms within LLMs may remain susceptible to adversarial prompting or so-called ”jailbreaking”, raising additional concerns regarding patient safety[46].

Overall, the current literature indicates that GenAI may serve a supportive role in mental healthcare, particularly as an adjunctive tool rather than a replacement for mental health professionals[29,37]. Potential applications are especially relevant in low-intensity support, psychoeducation, administrative optimisation, and preliminary symptom monitoring. However, safe and effective integration of these technologies into clinical practice will require ongoing human oversight, robust ethical and regulatory frameworks, and further development of clinically specialised models.

Future research should prioritise large-scale, real-world studies evaluating effectiveness, safety, acceptability, and long-term clinical impact across diverse populations and healthcare systems. Greater standardisation of evaluation methods and direct comparison between different models may improve the quality and interpretability of future evidence. As GenAI technologies continue to evolve rapidly, it is essential to ensure that their development remains aligned with clinical standards, ethical principles, and patient-centred care.

References

[1] Lekadir K, Quaglio G, Tselioudis Garmendia A, Gallin C. Artificial intelligence in healthcare: applications, risks, and ethical and societal impacts. Brussels: European Parliamentary Research Service; 2022.

[2] Sun J, Lu T, Shao X, et al. Practical AI application in psychiatry: historical review and future directions. Mol Psychiatry. 2025. Epub ahead of print. doi:10.1038/s41380-025-03072-3.

[3] D'Souza RF, Amanullah S, Mathew M, Surapaneni KM. Appraising the performance of ChatGPT in psychiatry using 100 clinical case vignettes. Asian J Psychiatr. 2023;89:103770. doi:10.1016/j.ajp.2023.103770.

[4] Kolding S, Lundin RM, Hansen L, Ostergaard SD. Use of generative artificial intelligence (AI) in psychiatry and mental health care: a systematic review. Acta Neuropsychiatr. 2024;37. doi:10.1017/neu.2024.50.

[5] Peng C, Yang X, Chen A, et al. A study of generative large language model for medical research and healthcare. NPJ Digit Med. 2023;6(1). doi:10.1038/s41746-023-00958-w.

[6] Dergaa I, Fekih-Romdhane F, Hallit S, et al. ChatGPT is not ready yet for use in providing mental health assessment and interventions. Front Psychiatry. 2023;14:1277756. doi:10.3389/fpsyt.2023.1277756.

[7] Tortora L. Beyond discrimination: generative AI applications and ethical challenges in forensic psychiatry. Front Psychiatry. 2024;15:1346059. doi:10.3389/fpsyt.2024.1346059.

[8] Le Glaz A, Haralambous Y, Kim-Dufor DH, et al. Machine learning and natural language processing in mental health: systematic review. J Med Internet Res. 2021;23(5):e15708. doi:10.2196/15708.

[9] 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. doi:10.2196/57400.

[10] Crema C, Attardi G, Sartiano D, Redolfi A. Natural language processing in clinical neuroscience and psychiatry: a review. Front Psychiatry. 2022;13:946387. doi:10.3389/fpsyt.2022.946387.

[11] Google Cloud [Internet]. Mountain View: Google Cloud; MedLM; 2024. Disponível em: https://cloud.google.com/vertex-ai/generative-ai/docs/medlm/overview. Consultado pela última vez a 2026/05/24.

[12] Acharya DB, Kuppan K, Divya B. Agentic AI: autonomous intelligence for complex goals - a comprehensive survey. IEEE Access. 2025;13:18912-18936. doi:10.1109/ACCESS.2025.3532853.

[13] Bednarczyk L, Reichenpfader D, Gaudet-Blavignac C, et al. Scientific evidence for clinical text summarization using large language models: scoping review. J Med Internet Res. 2025;27:e68998. doi:10.2196/68998.

[14] De Choudhury M, Pendse SR, Kumar N. Benefits and harms of large language models in digital mental health [Internet]. Ithaca: arXiv; 2023. Disponível em: http://arxiv.org/abs/2311.14693. Consultado pela última vez a 2026/05/24.

[15] Haq IU, Chhatwal K, Sanaka K, Xu B. Artificial intelligence in cardiovascular medicine: current insights and future prospects. Vasc Health Risk Manag. 2022;18:517-528. doi:10.2147/VHRM.S279337.

[16] Merabet A, Saighi A, Saad H, et al. AI for colon cancer: a focus on classification, detection, and predictive modeling. Int J Med Inform. 2025;206:106115. doi:10.1016/j.ijmedinf.2025.106115.

[17] Hussen BM, Abdullah SR, Hidayat HJ, Samsami M, Taheri M. Integrating AI and RNA biomarkers in cancer: advances in diagnostics and targeted therapies. Cell Commun Signal. 2025;23(1):430. doi:10.1186/s12964-025-02434-2.

[18] Zhou XY, Guo Y, Shen M, Yang GZ. Application of artificial intelligence in surgery. Front Med. 2020;14(4):417-430. doi:10.1007/s11684-020-0770-0.

[19] Moglia A, Georgiou K, Georgiou E, Satava RM, Cuschieri A. A systematic review on artificial intelligence in robot-assisted surgery. Int J Surg. 2021;95:106151. doi:10.1016/j.ijsu.2021.106151.

[20] Tong ACY, Wong KTY, Chung WWT, Mak WWS. Effectiveness of topic-based chatbots on mental health self-care and mental well-being: randomized controlled trial. J Med Internet Res. 2025;27:e70436. doi:10.2196/70436.

[21] Zhong W, Luo J, Zhang H. The therapeutic effectiveness of artificial intelligence-based chatbots in alleviation of depressive and anxiety symptoms in short-course treatments: a systematic review and meta-analysis. J Affect Disord. 2024;356:459-469. doi:10.1016/j.jad.2024.04.057.

[22] Fitzpatrick KK, Darcy A, Vierhile M. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. JMIR Ment Health. 2017;4(2):e19. doi:10.2196/mental.7785.

[23] Dehbozorgi R, Zangeneh S, Khooshab E, et al. The application of artificial intelligence in the field of mental health: a systematic review. BMC Psychiatry. 2025;25(1). doi:10.1186/s12888-025-06483-2.

[24] Elyoseph Z, Hadar-Shoval D, Asraf K, Lvovsky M. ChatGPT outperforms humans in emotional awareness evaluations. Front Psychol. 2023;14:1199058. doi:10.3389/fpsyg.2023.1199058.

[25] Blease C, Torous J, McMillan B, Hägglund M, Mandl KD. Generative language models and open notes: exploring the promise and limitations. JMIR Med Educ. 2024;10:e51183. doi:10.2196/51183.

[26] Chen C, Lam KT, Yip KM, et al. Comparison of an AI chatbot with a nurse hotline in reducing anxiety and depression levels in the general population: pilot randomized controlled trial. JMIR Hum Factors. 2025;12:e65785. doi:10.2196/65785.

[27] Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. NPJ Digit Med. 2023;6(1). doi:10.1038/s41746-023-00873-0.

[28] Wang Y, Li X, Zhang Q, Yeung D, Wu Y. Effect of a cognitive behavioral therapy-based AI chatbot on depression and loneliness in Chinese university students: randomized controlled trial with financial stress moderation. JMIR Mhealth Uhealth. 2025;13:e63806. doi:10.2196/63806.

[29] Alanezi F. Assessing the effectiveness of ChatGPT in delivering mental health support: a qualitative study. J Multidiscip Healthc. 2024;17:461-471. doi:10.2147/JMDH.S447368.

[30] Hristidis V, Ruggiano N, Brown EL, Ganta SRR, Stewart S. ChatGPT vs Google for queries related to dementia and other cognitive decline: comparison of results. J Med Internet Res. 2023;25:e48966. doi:10.2196/48966.

[31] Amin S, Kawamoto CT, Pokhrel P. Exploring the ChatGPT platform with scenario-specific prompts for vaping cessation. Tob Control. 2025;34(2):251-253. doi:10.1136/tc-2023-058009.

[32] Sabour S, Zhang W, Xiao X, et al. A chatbot for mental health support: exploring the impact of Emohaa on reducing mental distress in China. Front Digit Health. 2023;5:1133987. doi:10.3389/fdgth.2023.1133987.

[33] Tang Y, Chen L, Chen Z, et al. EmoEden: applying generative artificial intelligence to emotional learning for children with high-function autism. In: Proceedings of the CHI Conference on Human Factors in Computing Systems. New York: Association for Computing Machinery; 2024. doi:10.1145/3613904.3642899.

[34] Koegel LK, Ponder E, Bruzzese T, et al. Using artificial intelligence to improve empathetic statements in autistic adolescents and adults: a randomized clinical trial. J Autism Dev Disord. 2025. Epub ahead of print. doi:10.1007/s10803-025-06734-x.

[35] Deng C, Lai S, Zhou C, et al. ASD-Chat: an innovative dialogue intervention system for children with autism based on LLM and VB-MAPP [Internet]. Ithaca: arXiv; 2024. Disponível em: http://arxiv.org/abs/2409.01867. Consultado pela última vez a 2026/05/24.

[36] McFayden TC, Bristol S, Putnam O, Harrop C. ChatGPT: artificial intelligence as a potential tool for parents seeking information about autism. Cyberpsychol Behav Soc Netw. 2024;27(2):135-148. doi:10.1089/cyber.2023.0202.

[37] Blease C, Worthen A, Torous J. Psychiatrists' experiences and opinions of generative artificial intelligence in mental healthcare: an online mixed methods survey. Psychiatry Res. 2024;333:115724. doi:10.1016/j.psychres.2024.115724.

[38] Pombal J, D'Eon M, Guerreiro NM, Martins PH, Farinhas A, Rei R. MindEval: benchmarking language models on multi-turn mental health support [Internet]. Ithaca: arXiv; 2025. Disponível em: http://arxiv.org/abs/2511.18491. Consultado pela última vez a 2026/05/24.

[39] Herrmann-Werner A, Festl-Wietek T, Holderried F, et al. Assessing ChatGPT's mastery of Bloom's taxonomy using psychosomatic medicine exam questions: mixed-methods study. J Med Internet Res. 2024;26:e52113. doi:10.2196/52113.

[40] Hadar-Shoval D, Elyoseph Z, Lvovsky M. The plasticity of ChatGPT's mentalizing abilities: personalization for personality structures. Front Psychiatry. 2023;14:1234397. doi:10.3389/fpsyt.2023.1234397.

[41] He W, Zhang W, Jin Y, Zhou Q, Zhang H, Xia Q. Physician versus large language model chatbot responses to web-based questions from autistic patients in Chinese: cross-sectional comparative analysis. J Med Internet Res. 2024;26:e54706. doi:10.2196/54706.

[42] Sezgin E, Chekeni F, Lee J, Keim S. Clinical accuracy of large language models and Google Search responses to postpartum depression questions: cross-sectional study. J Med Internet Res. 2023;25:e49240. doi:10.2196/49240.

[43] Hwang G, Lee DY, Seol S, et al. Assessing the potential of ChatGPT for psychodynamic formulations in psychiatry: an exploratory study. Psychiatry Res. 2024;331:115655. doi:10.1016/j.psychres.2023.115655.

[44] Elyoseph Z, Levkovich I, Shinan-Altman S. Assessing prognosis in depression: comparing perspectives of AI models, mental health professionals and the general public. Fam Med Community Health. 2024;12(Suppl 1):e002583. doi:10.1136/fmch-2023-002583.

[45] Lee QY, Chen M, Ong CW, Ho CSH. The role of generative artificial intelligence in psychiatric education: a scoping review. BMC Med Educ. 2025;25(1). doi:10.1186/s12909-025-07026-9.

[46] Schoene AM, Canca C. For argument's sake, show me how to harm myself!: jailbreaking LLMs in suicide and self-harm contexts [Internet]. Ithaca: arXiv; 2025. Disponível em: http://arxiv.org/abs/2507.02990. Consultado pela última vez a 2026/05/24.

[47] Bouguettaya A, Team V, Stuart EM, Aboujaoude E. AI-driven report-generation tools in mental healthcare: a review of commercial tools. Gen Hosp Psychiatry. 2025;94:150-158. doi:10.1016/j.genhosppsych.2025.02.018.

Tables

Table I. Domain-Specific Language Models in Healthcare AI

Model	Description
BERT (Bidirectional Encoder Representations from Transformers)	A foundational language model developed by Google, pre-trained on large general-purpose corpora such as Wikipedia and BookCorpus. It introduced bidirectional contextual understanding through techniques such as masked language modelling and next sentence prediction, serving as the baseline architecture for subsequent domain-specific adaptations.
BioBERT	A domain-adapted version of BERT pre-trained on large-scale biomedical corpora, including PubMed abstracts and full-text articles, designed to improve performance in biomedical text mining and information extraction tasks.
Clinical (Bio)BERT	A further specialised model derived from BERT and BioBERT, fine-tuned on clinical corpora such as EHR. This targeted adaptation enables superior performance in a range of biomedical NLP tasks compared to general and biomedical pre-trained models.
UMLS-BERT	An extension of BERT-based models that incorporates structured expert knowledge from the UMLS during pre-training, resulting in more clinically informative representations and improved performance in tasks such as NER and clinical inference.
MedLM	A large-scale clinical language model that builds upon previous approaches by leveraging extensive healthcare datasets and advanced fine-tuning techniques, designed to support complex clinical applications including medical question answering and clinical documentation summarisation.

Table II. Distribution of included studies according to clinical area and intended GenAI application

Clinical area / target condition	Symptom Assessment and Triage	Patient Support	Clinical Decision Support	Administrative Support	Evaluation / Benchmarking of AI Systems
General psychiatry / mixed psychiatric conditions	D'Souza et al., 2023 – ChatGPT 3.5	Alanezi, 2024 – ChatGPT	–	Blease et al., 2024 – LLM-based chatbots	Pombal et al., 2025 – MindEval / LLM benchmarking framework
Depression, anxiety, loneliness and psychological distress	–	Chen et al., 2025 – AI chatbot; Wang et al., 2025 – CBT-based AI chatbot; Sabour et al., 2023 – Emohaa	Elyoseph et al., 2024 – LLMs for depression prognosis	–	–
Insomnia / sleep difficulties	Dergaa et al., 2024 – ChatGPT	Sabour et al., 2023 – Emohaa	–	–	–
Autism spectrum disorder / high-functioning autism	–	McFayden et al., 2024 – ChatGPT-4; Deng et al., 2024 – ASD-Chat; Tang et al., 2024 – EmoEden; Koegel et al., 2025 – Noora	–	–	He et al., 2024 – ChatGPT-4 and ERNIE Bot
Dementia and cognitive decline	–	Hristidis et al., 2023 – ChatGPT vs Google	–	–	–
Postpartum depression	–	–	–	–	Sezgin et al., 2023 – ChatGPT-4, Bard and Google Search
Personality disorders / mentalising abilities	–	–	–	–	Hadar-Shoval et al., 2023 – ChatGPT
Emotional awareness / transdiagnostic emotional processes	–	–	–	–	-
Psychosomatic medicine	–	–	–	–	Herrmann-Werner et al., 2024 – GPT-4
Psychodynamic case formulation	–	–	–	–	Hwang et al., 2024 – ChatGPT / GPT-4
Substance use / vaping cessation	–	Amin et al., 2023 – ChatGPT	–	–	–

Table III. Characteristics of the included studies

Authors	Publishing year	DOI	Study design	Area(s)	Technology	Disease/Condition	Main Results
Amin S, et al.	2023	10.1136/tc-2023-058009	Simulation study	Patient Support	ChatGPT (GPT-3.5)	Vaping cessation -substance use behavior	ChatGPT generated responses that were generally appropriate, coherent, and aligned with users' needs, receiving high ratings across multiple quality domains; however, limitations were noted in handling complex or urgent situations and in the consistency of some recommendations.
Franco D'Souza R, et al.	2023	10.1016/j.ajp.2023.103770	Vignette-based simulation study	Symptom Assessment and Triage	ChatGPT 3.5	100 different psychiatric illnesses	ChatGPT responses were predominantly rated in the highest categories (Grade A or B: 92%) across diagnostic and management tasks, based on evaluation of the psychiatric case vignettes.
Hadar-Shoval D., et al.	2023	10.3389/fpsyt.2023.1234397	Experimental simulation study	Evaluation / Benchmarking of AI Systems	ChatGPT (GPT-3.5)	Borderline personality disorder and Schizoid personality disorder	ChatGPT generated responses that differentiated between personality profiles, producing more complex and intense emotional descriptions for borderline presentations and more reduced emotional patterns for schizoid traits, with significantly higher emotional awareness scores in the former condition.
Hristidis V, et al.	2023	10.2196/48966	Comparative evaluation study	Patient Support	ChatGPT vs Google	Dementia and cognitive decline	ChatGPT produced answers that better matched users' questions and were generally more neutral, whereas Google provided more up-to-date information and more clearly sourced content; both platforms showed limitations in readability for lay users.

Table III. Characteristics of the included studies (continued)

Authors	Publishing year	DOI	Study design	Area(s)	Technology	Disease/Condition	Main Results
Sabour S, et al.	2023	10.3389/fdgth.2023.1133987	Randomized controlled trial	Patient Support	Emohaa (CBT-based and emotional support chatbot)	Mental distress - depression, negative affect, insomnia	Participants in the intervention groups showed greater reductions in depression, negative affect, and insomnia compared to the control group, with statistically significant differences between groups; no consistent differences were observed for anxiety.
Sezgin E., et al.	2023	10.2196/49240	Cross-sectional comparative study	Evaluation / Benchmarking of AI Systems	ChatGPT-4, Bard, and Google Search	Postpartum depression	ChatGPT achieved the highest clinical accuracy scores compared to Bard and Google Search, with consistently more complete and appropriate responses, while the other platforms showed lower performance and greater variability in answer quality.
Alanezi F, et al.	2024	10.2147/JMDH.S447368	Qualitative study	Patient Support	ChatGPT	Various mental health conditions	The chatbot provided informational resources and general guidance but did not demonstrate diagnostic capabilities or personalised interaction comparable to human therapists.
Blease C., et al.	2024	10.1016/j.psychres.2024.115724	Mixed-methods survey study	Administrative Support	LLM-based chatbots (ChatGPT, Google Bard, Bing AI)	Mental healthcare - general psychiatry	Psychiatrists reported mixed views on the use of AI tools, identifying benefits mainly in administrative tasks such as documentation, while raising concerns about accuracy, bias, and patient safety; a majority also anticipated increased patient use of these tools and highlighted the need for further training.
Deng C., et al.	2024	arXiv:2409.01867	Experimental comparative study	Patient Support	ASD-Chat (LLM-based dialogue system grounded in VB-MAPP)	Autism spectrum disorder (children)	Children showed higher engagement with the AI system, while response quality was slightly lower compared to human intervention; behavioural and neurophysiological measures were otherwise similar across conditions.

Table III. Characteristics of the included studies (continued)

Authors	Publishing year	DOI	Study design	Area(s)	Technology	Disease/Condition	Main Results
Dergaa I, et al.	2024	10.3389/fpsyt.2023.1277756	Simulation study	Symptom Assessment and Triage	ChatGPT	Insomnia and sleep difficulties	Qualitative clinician assessment showed adequate but nonspecific responses in simple cases, with clinically inadequate outputs in complex scenarios due to lack of follow-up and non-tailored recommendations; no quantitative metrics reported.
Elyoseph Z, et al.	2024	10.1136/fmch-2023-002583	Vignette-based comparative study	Clinical Decision Support	Large language models (ChatGPT-3.5, GPT-4, Claude.AI, Bard)	Depression	Most AI models aligned with healthcare professionals on predicting clinical prognosis and long-term outcomes. However, different versions of ChatGPT produced more negative prognoses or identified more negative projected consequences.
He W., et al.	2024	10.2196/54706	Cross-sectional comparative study	Evaluation / Benchmarking of AI Systems	ChatGPT-4 and ERNIE Bot	Autism spectrum disorder	ChatGPT showed performance similar to physicians in relevance and accuracy and higher empathy scores, while physicians were more often preferred and rated higher in usefulness; ERNIE Bot generally showed lower performance across most domains.
Herrmann-Werner A., et al.	2024	10.2196/52113	Mixed-methods study (quantitative + qualitative analysis)	Evaluation / Benchmarking of AI Systems	GPT-4	Psychosomatic medicine	GPT-4 achieved high accuracy in answering multiple-choice questions, with performance consistently above passing thresholds; errors were mainly related to gaps in factual recall and basic understanding, while more complex reasoning tasks were less frequently affected.

Table III. Characteristics of the included studies (continued)

Authors	Publishing year	DOI	Study design	Area(s)	Technology	Disease/Condition	Main Results
Hwang G., et al.	2024	10.1016/j.psychres.2023.115655	Exploratory evaluation study	Evaluation / Benchmarking of AI Systems	ChatGPT (GPT-4)	Psychodynamic case formulation - psychiatric case	Outputs were rated as appropriate when based on full case descriptions or psychodynamic concepts, whereas performance was lower when prompts relied on keywords alone; responses also adapted to different psychoanalytic models with generally acceptable quality.
McFayden T.C., et al.	2024	10.1089/cyber.2023.0202	Qualitative evaluation study	Patient Support	ChatGPT-4	Autism spectrum disorder	Responses were generally accurate, clear, and easy to understand, with high scores across quality domains; however, guidance for actionable next steps was limited and many provided references or links were inaccurate or non-functional.
Tang Y, et al.	2024	10.1145/3613904.3642899	Pre-post intervention study	Patient Support	EmoEden (generative AI-based system using LLM + text-to-image)	High-functioning autism (children)	Use of the system was associated with improvements in emotional recognition and expression across participants, with gains observed after the intervention period and maintained in short-term follow-up assessments.
Chen C, et al.	2025	10.2196/65785	Pilot randomised controlled trial	Patient Support	AI Chatbot vs. Nurse Hotline	Anxiety and depression	The AI chatbot produced a significant reduction in anxiety and depression scores.
Koegel L.K., et al.	2025	10.1007/s10803-025-06734-x	Randomized controlled trial (waitlist control)	Patient Support	Noora (LLM-based conversational training system)	Autism spectrum disorder (adolescents and adults)	Participants using the intervention showed greater improvement in empathetic responses during structured tasks and real-life conversation compared to the control group, with statistically significant between-group differences; most participants also improved or maintained performance over time.

Table III. Characteristics of the included studies (continued)

Authors	Publishing year	DOI	Study design	Area(s)	Technology	Disease/Condition	Main Results
Pombal J, et al.	2025	10.48550/arXiv.2511.18491	Benchmarking study	Evaluation / Benchmarking of AI Systems	LLMs (Benchmarking framework)	General mental health support	Developed a framework to benchmark LLMs on multi-turn mental health support using simulated patients; the LLM judge correlated strongly with human psychologists
Wang Y, et al.	2025	10.2196/63806	Randomised controlled trial	Patient Support	CBT-Based AI Chatbot (Psy-Bot)	Depression and loneliness	The culturally adapted AI chatbot helped university students relieve anxiety and loneliness by delivering automated CBT exercises.

Figures

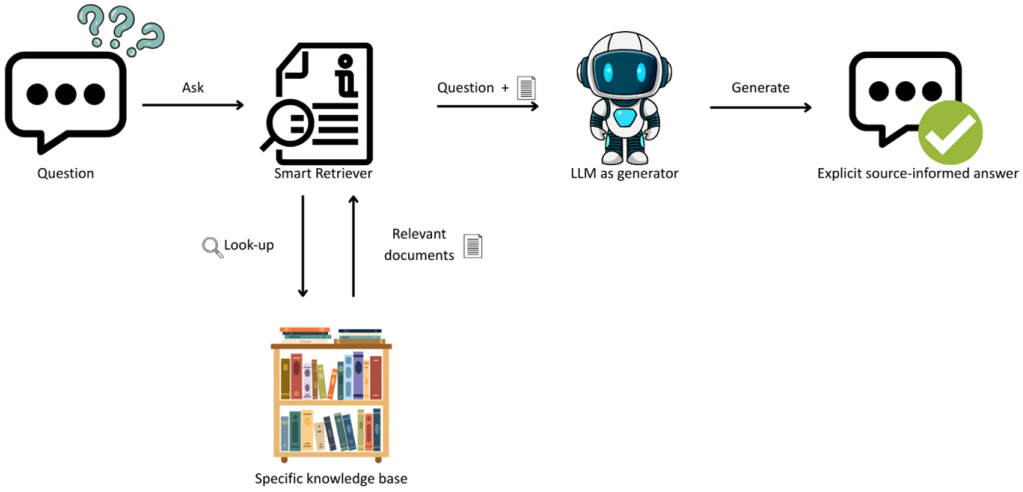


Fig. 1. Retrieval-Augmented Generation (RAG).

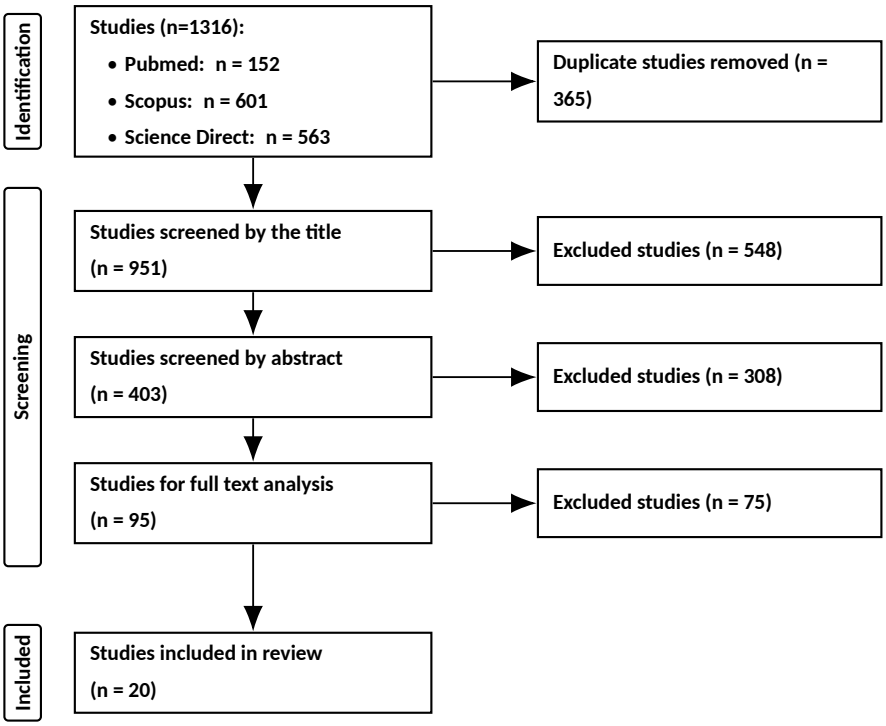


Fig. 2. Flowchart of the study selection process.

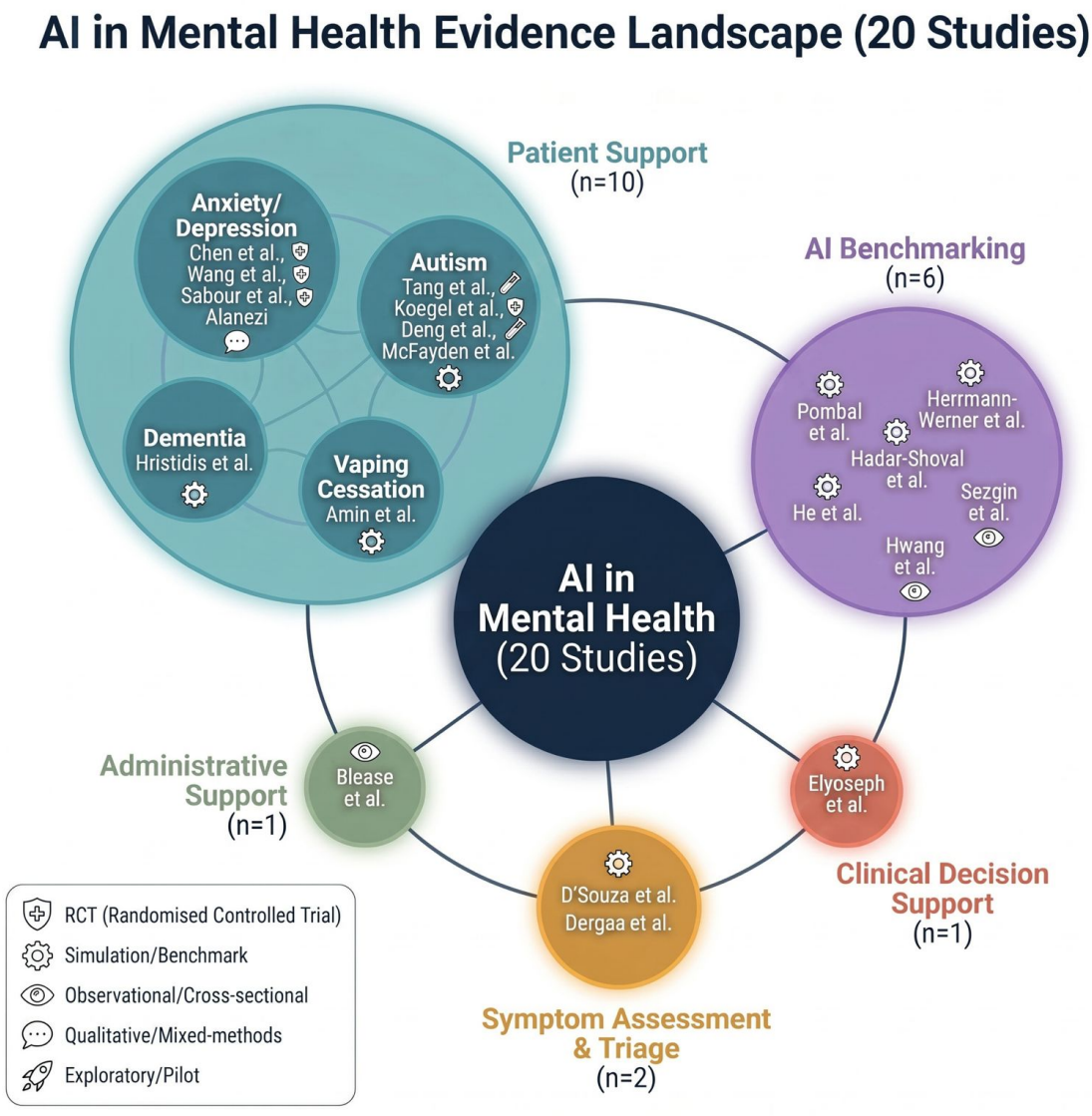


Fig. 3. AI in Mental Health Evidence Landscape (20 Studies).

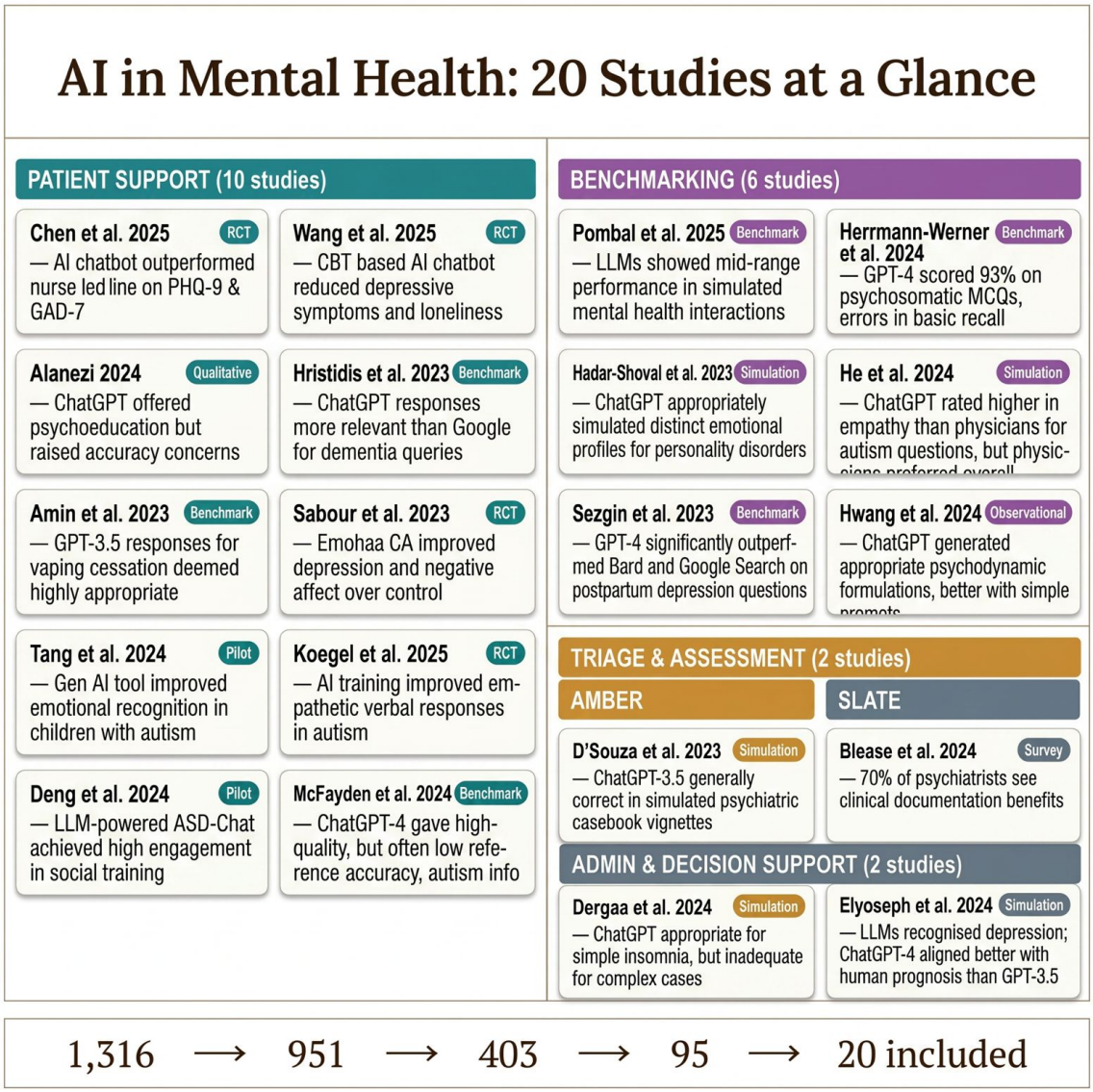


Fig. 4. AI in Mental Health: 20 Studies at a Glance.

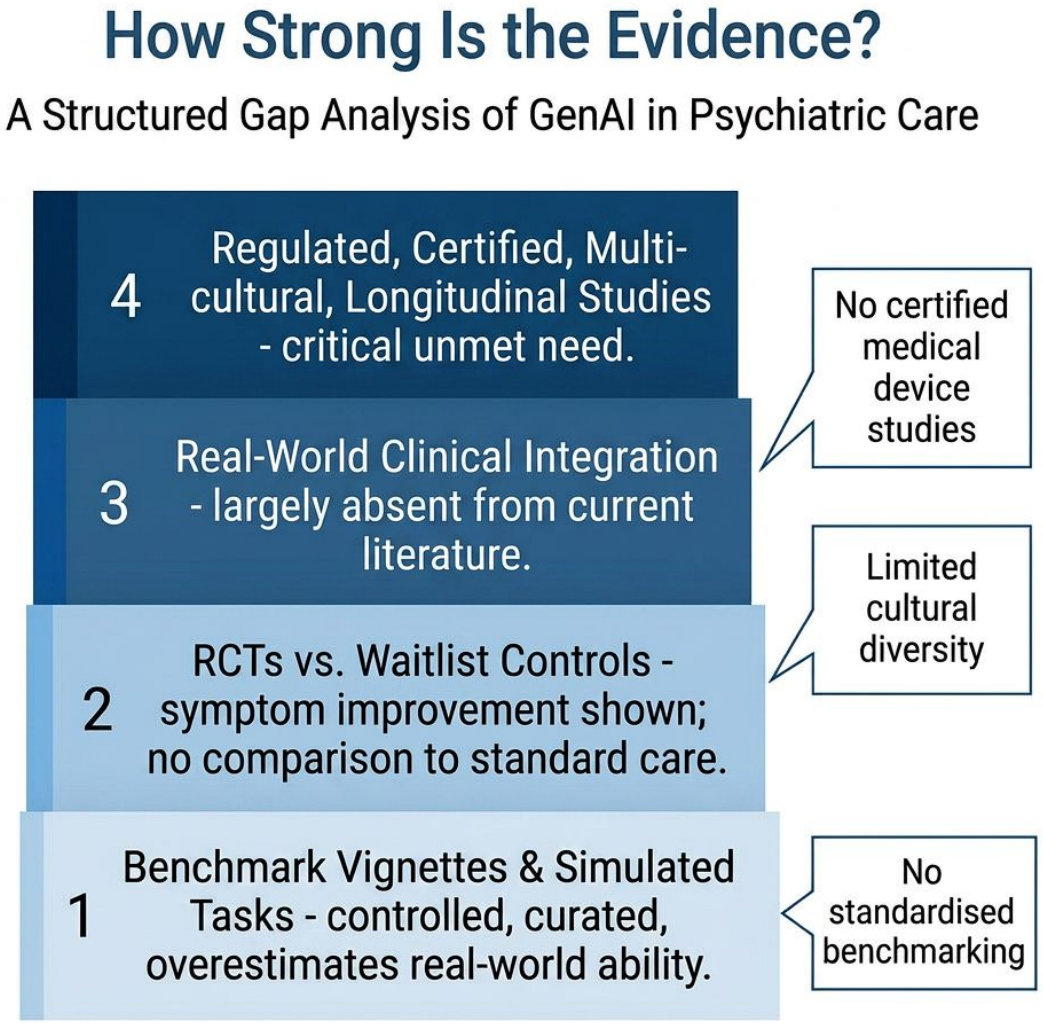


Fig. 5. Strength of the evidence and evidence gaps in GenAI psychiatric care.

INSTITUTO DE **CIÊNCIAS BIOMÉDICAS ABEL SALAZAR**