Review

Not peer-reviewed version

Large Language Models for Cardiovascular Disease, Cancer, and Mental Disorders: A Review of Systematic Reviews

Andreas Triantafyllidis * , Sofia Segkouli , Stelios Kokkas , Anastasios Alexiadis , Eirini Lithoxoidou , George Manias , Athos Antoniades , Konstantinos Votis , Dimitrios Tzovaras

Posted Date: 31 October 2025

doi: 10.20944/preprints202510.2480.v1

Keywords: large language models; generative AI; digital health; literature review

 doi:10.20944/preprints202510.2480.v1

Review

Large Language Models for Cardiovascular Disease, Cancer, and Mental Disorders: A Review of Systematic Reviews

Andreas Triantafyllidis [1,*], Sofia Segkouli [1], Stelios Kokkas [1], Anastasios Alexiadis [1], Eirini Lithoxoidou [1], George Manias [2], Athos Antoniades [3], Konstantinos Votis [1] and Dimitrios Tzovaras [1]

[1] Information Technologies Institute, Centre for Research and Technology Hellas, Thessaloniki 57001, Greece
[2] Department of Digital Systems, University of Piraeus, Piraeus, Greece
[3] Stremble Ventures Ltd., Limassol, Cyprus
* Correspondence: atriand@iti.gr; Tel.: +30-2311-257701

Abstract

Background/Objective: The use of Large Language Models (LLMs) has recently gained significant interest from the research community toward the development and adoption of Generative Artificial Intelligence (GenAI) solutions for healthcare. The present work introduces the first meta-review (i.e., review of systematic reviews) in the field of LLMs for chronic diseases, focusing particularly on cardiovascular, cancer, and mental diseases, with the goal to identify their value in patient care, as well as challenges for their implementation and clinical application. **Methods:** A literature search in the bibliographic databases of PubMed and Scopus was conducted following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines, in order to identify systematic reviews incorporating LLMs. The original studies included in the reviews were synthesized according to their target disease, specific application, used LLMs, data sources, accuracy, and key outcomes. **Results:** The literature search identified 5 systematic reviews respecting our inclusion and exclusion criteria, which examined 81 unique LLM-based solutions. The highest percentage of the solutions targeted mental disease (70 studies, 86%), followed by cancer (6 studies, 7%) and cardiovascular disease (5 studies, 6%). Generative Pre-trained Transformer (GPT) models were used in most studies (45 studies, 55%), followed by Bidirectional Encoder Representations from Transformers (BERT) models (33 studies, 40%). Key application areas included depression detection and classification (31 studies, 38%), suicidal ideation detection (6 studies, 7%), question answering based on treatment guidelines and recommendations (6 studies, 7%), and emotion classification (4 studies, 5%). Studies were highly heterogeneous, with most studies (41 studies, 50%) focusing on clinical/diagnostic accuracy using metrics for correct diagnosis, agreement with guidelines or model performance, whereas other studies (34 studies, 42%) were descriptive and focused on narrative outcomes, usability, trust or plausibility. The most significant found challenges in the development and evaluation of LLMs include inconsistent accuracy, bias detection and mitigation, model transparency, data privacy, need for continual human oversight, ethical concerns and guidelines, as well as the design and conduction of high-quality studies. **Conclusion:** The results of this review suggest LLMs could become valuable tools for enhancing diagnostic precision and decision support in cardiovascular disease, cancer and mental health. Given the limited number of studies included in this review and their moderate quality, we urge the research community to conduct more investigations in real-world intervention settings to better demonstrate the clinical utility of LLMs.

Keywords: large language models; generative AI; digital health; literature review

 doi:10.20944/preprints202510.2480.v1

1. Introduction

The advent of Large Language Models (LLMs), exemplified by platforms such as ChatGPT, represents a significant advancement in Generative Artificial Intelligence (GenAI) with far-reaching implications for improving healthcare [1]. LLMs are trained on massive amounts of data and are able to generate human-like text, answer questions, and complete other language-related tasks with high accuracy [2]. LLMs have demonstrated strong capacity for complex clinical reasoning, patient education, and decision support across a variety of medical fields such as psychiatry [3], cardiology [4], and oncology [5].

The adoption of LLMs in healthcare emerges as a promising step to provide more efficient, safer, and personalised care for patients, mainly because of their capability for swift natural language communication with the users, along with synthesis, summarisation, and contextual reasoning over diverse clinical information, around the clock. The acceptability, usability, and potential of LLMs in improving health and well-being have been explored in a number of previous studies [6–8].

As the body of peer-reviewed research on LLMs in healthcare expands rapidly, there is a pressing need to establish a solid evidence base regarding their clinical effectiveness and practical value. To address this gap, we present a review of systematic reviews examining the use of LLMs among patients with chronic diseases—with particular emphasis on cardiovascular, oncological, and mental health conditions due to their significant global burden. To the authors' knowledge, this meta-review is the first in the field of LLMs for healthcare.

Our work aims to analyse the features, outcomes, and methodological characteristics of LLM-based interventions and tools, while synthesizing recent evidence and identifying key challenges in this fast-evolving domain. Unlike disease-specific analyses, the current review adopts a cross-condition perspective, enabling a broader understanding of how LLMs are being leveraged across diverse chronic care contexts. By systematically mapping the current state of applications, evaluating their effectiveness, and highlighting limitations, this study seeks to inform researchers, developers, clinicians, and policymakers in designing and implementing more robust and impactful LLM-driven health interventions, toward realising the full potential of GenAI application in the clinical practice.

2. Methodology

We searched the bibliographic databases of PubMed and Scopus to identify systematic reviews of the application of LLMs for cardiovascular disease, cancer, and mental disorders, as reported in manuscripts published after 2022, the year in which the interest in LLMs was exploded due to the appearance of ChatGPT. The inclusion criteria were: a) the review should be defined as systematic, focus on the effectiveness of the LLMs, and follow reporting guidelines such as the Preferred Reporting Items for Systematic review and Meta-Analyses (PRISMA) [9], or the Cochrane guidelines [10], b) the review should report LLM-based tools or interventions targeted at chronically ill individuals diagnosed with cardiovascular disease, cancer, or mental disorders, c) the paper should be written in English. We used the keyword query ("LLM" OR "large language model" OR "chatbot" OR "conversational" OR "bot" OR "digital assistant" OR "virtual assistant" OR "digital agent" OR "virtual agent") for search within the title, abstract and keywords of the manuscripts, and restricted the search to a review type of articles. Reviews examining interventions or tools which were not focused on leveraging LLMs, were excluded. Furthermore, reviews focusing exclusively on other medical fields such as surgery or medical education were excluded. Reviews not examining quantitative outcomes, surveys, and protocol papers were also excluded from the review.

The selection and review of papers were conducted independently by four reviewers (authors AT, SS, AA, SK) to ensure relevance based on predefined inclusion and exclusion criteria and to minimize potential selection bias or errors. Following the literature search, both abstracts and full-text manuscripts were screened. Studies that did not meet the inclusion criteria were excluded, and only those for which consensus among all reviewers was achieved were retained.

 doi:10.20944/preprints202510.2480.v1

The methodological quality of the included systematic reviews was evaluated using the second version of the AMSTAR tool (A Measurement Tool to Assess Systematic Reviews), which has demonstrated reliability [11]. Data from the primary studies included within these reviews were synthesized (by AT) according to several dimensions: Target disease, application area, LLM used, leveraged data sources, model accuracy, and key outcomes.

3. Results

3.1. Literature Search Outcomes

Our literature search in the PubMed and Scopus databases was conducted on November, 2024 with the last search update taking place on May, 2025. The search yielded 422 records from the Scopus database, and 120 records from Pubmed. After removing all duplicates in the Mendeley© bibliography management software [12], and applying our inclusion and exclusion criteria, 24 articles remained for full manuscript reading. Finally, 5 papers (systematic reviews) were included [13–17]. Reasons for paper exclusion are shown in Figure 1.

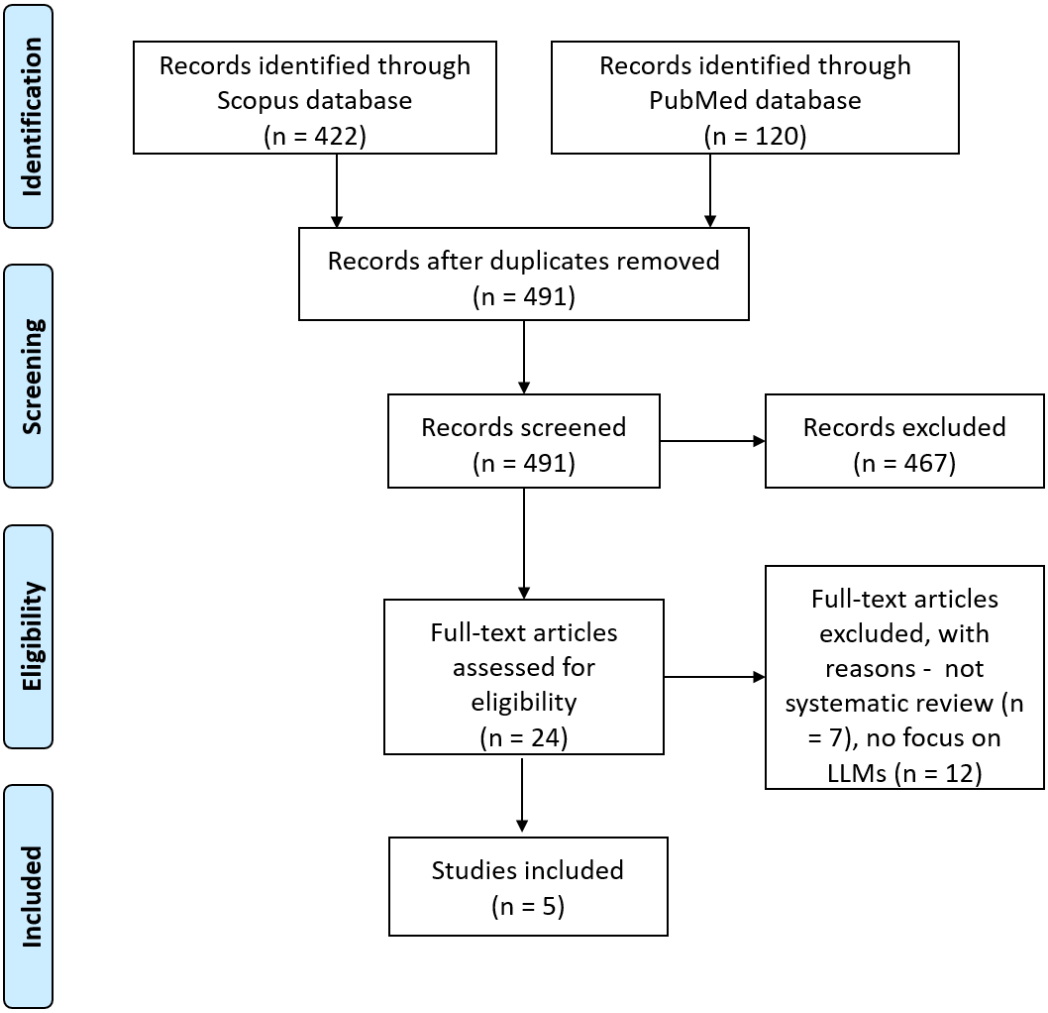

Figure 1. PRISMA flow diagram for study inclusion.

3.2. Quality Assessment of Reviews and Original Studies

The quality of the included review studies as assessed through AMSTAR 2 criteria can be seen in Table 1. All reviews performed study selection in duplicate and discussed the heterogeneity of studies and results. However, the studies did not explicitly include the components of Population, Intervention, Comparison, Outcome (PICO), did not explain the selection of study designs for

 doi:10.20944/preprints202510.2480.v1

inclusion in the reviews, did not perform a comprehensive literature search (searches of references, registries, grey literature, consultation with experts), and did not take into account the risk of bias in individual studies when interpreting the results. Three reviews [14,16,17], assessed the risk of bias of the individual studies using tools such as Quality Assessment of Diagnostic Accuracy Studies (QUADAS 2), Risk of Bias 2 Tool, Risk Of Bias In Non-randomised Studies - of Interventions (ROBINS-I), and Prediction model Risk Of Bias Assessment Tool (PROBAST). The individual studies were reported to be of high risk of bias except one in the review for breast cancer management by Sorin et al. [14], while Guo et al. [16] have reported overall a low risk of bias in the studies for mental health applications. Omar and Levkovich [17] in their review focusing on depression, reported an overall mixed picture in terms of risk of bias assessment. Although the majority of studies were reported to be of low risk of bias in measurements and outcomes, several studies presented moderate biases due to confounding factors and participant selection that may have affected the applicability of the results.

Table 1. Quality assessment of the included reviews according to AMSTAR 2 criteria (Y: Yes, N: No, PY: Partial Yes) of included studies (Original items concerning meta-analysis and quantitative synthesis were removed because they deemed to be out of scope).

AMSTAR 2 Criteria	Study				
	Sharma et al. [13]	**Sorin et al. [14]**	**Omar et al. [15]**	**Guo et al. [16]**	**Omar and Levkovich [17]**
1. Did the research questions and inclusion criteria for the review include the components of PICO?	N	N	N	N	N
2. Did the report of the review contain an explicit statement that the review methods were established prior to the conduct of the review and did the report justify any significant deviations from the protocol?	N	N	PY	PY	PY
3. Did the review authors explain their selection of the study designs for inclusion in the review?	N	N	N	N	N
4. Did the review authors use a comprehensive literature search strategy?	PY	N	PY	PY	PY
5. Did the review authors perform study selection in duplicate?	Y	Y	Y	Y	Y
6. Did the review authors perform data extraction in duplicate?	Y	N	Y	Y	Y
7. Did the review authors provide a list of excluded studies and justify the exclusions?	N	PY	N	Y	PY
8. Did the review authors describe the included studies in adequate detail?	N	N	PY	PY	PY
9. Did the review authors use a satisfactory technique for assessing the risk of bias (RoB) in individual studies that were included in the review?	N	Y	N	Y	Y
10. Did the review authors report on the sources of funding for the studies included in the review?	N	N	N	N	N

 doi:10.20944/preprints202510.2480.v1

11. Did the review authors account for RoB in individual studies when interpreting/ discussing the results of the review?	N	N	N	N	N
12. Did the review authors provide a satisfactory explanation for, and discussion of, any heterogeneity observed in the results of the review?	Y	Y	Y	Y	Y
13. Did the review authors report any potential sources of conflict of interest, including any funding they received for conducting the review?	Y	Y	Y	Y	Y

3.3. Characteristics of Individual Studies

In Table A1, the main characteristics of the 81 non-duplicate individual studies reported in the reviews are presented in terms of target disease, specific application, used LLMs, data sources, accuracy, and key outcomes. The majority of the solutions focused on mental health conditions (70 studies, 86%), followed by cancer (6 studies, 7%) and cardiovascular diseases (5 studies, 6%). Most studies employed Generative Pre-trained Transformer (GPT) models (45 studies, 55%), while Bidirectional Encoder Representations from Transformers (BERT) model and its variants were used in 33 studies (40%). The main application areas included depression detection and classification (31 studies, 38%), suicidal ideation detection (6 studies, 7%), question answering based on treatment guidelines and recommendations (6 studies, 7%), mental health intervention, e.g., support for loneliness (6 studies, 7%), and emotion classification (4 studies, 5%). The studies showed substantial heterogeneity: about half (41 studies, 50%) concentrated on clinical or diagnostic accuracy—evaluating correct diagnosis, guideline alignment, or model performance—while others (34 studies, 42%) were descriptive, focusing on aspects such as narrative outcomes, usability, trust, or plausibility.

3.4. Summary of LLMs Performance

The LLMs in depression applications (31 studies) achieved an accuracy in detecting the disease or its symptoms from 50% to 97% and an F1 score from 0.42 to 0.93. The studies reporting the highest performance used BERT, RoBERTa, AudiBERT, DistilBERT, and DeBERTa, and leveraged diverse datasets such as tweets, clinical interviews, or the mental health corpus database. The suicidal ideation studies (6 studies), utilised LLMs such as BERT, GPT, and XLNet, using social media posts (3 studies), as well as fictional patient data (3 studies). The reported achieved accuracy in 3 studies for suicidal ideation detection was from 87% to 95%. LLMs were found also to demonstrate good performance in several other mental health applications, such as classification of psychiatric conditions with BERT (F1 0.83) [18], mental health disorder prediction with BERT through Twitter (accuracy 97%) [19], and sentiment analysis through social media texts using GPT and Open Pre-trained Transformers (accuracy 86%) [20]. However, on other few occasions, LLMs did not demonstrate an acceptable performance, as in the case of the identification of diagnostic and management strategies in psychiatric conditions with GPT (accuracy 61%) [21]. All 6 studies reporting outcomes of diverse mental health interventions e.g., support for loneliness, mindfulness, and self-discovery, did not provide quantitative performance metrics, however they highlighted the usefulness of LLMs in mental health support, with only one study criticising the insufficiency of GPT in complex mental health scenarios.

In cardiology applications (5 studies), the LLMs were based on GPT and they were used for answering clinical questions. The evaluation of the performance was descriptive in the majority of the studies (3 studies), while an accuracy up to 64.5% was reported in the studies showing performance metrics (2 studies).

 doi:10.20944/preprints202510.2480.v1

In cancer applications (6 studies), GPT models were also used. 3 studies focused on tumor board clinical decision support, 2 studies on question-answering, and 1 study assessed the time and cost for developing LLM prompts. The accuracy of LLM responses compared with reviews of the tumor board was varied, reaching up to 70% with real patient data, and up to 95% with fictional patient data. The question-answering applications reached up to 98% accuracy. Regarding time and cost evaluation, LLM-based prompting was found to offer an efficient approach to extract key information from the medical records of breast cancer patients and to generate well-structured clinical datasets. This method is expected to significantly reduce the effort required in routine clinical practice and research.

3.5. Benefits of Using LLMs

3.5.1. Detection and Screening

LLMs have shown significant potential in the improvement of detection and screening for diseases. In mental health, transformer-trained LLMs have been applied to detect depression, suicidality, and anxiety, based on the examination of linguistic features and affectual cues within patient-written text or social media posts [15,16]. This strategy facilitates the early identification of mental illness within non-clinical environments, with the provision of scalable non-intrusive screening tools that can be used to monitor public health. The ability of LLMs to process large volumes of unstructured language data allows for timely detection of subtle indicators of distress that may escape conventional screening tools.

In oncology, benefits in screening procedures are also apparent. LLMs can extract tumor, receptor, and staging relevant variables from pathology as well as imaging reports with a high accuracy, reaching up to 98% [14]. This not only fastens the data-processing time but also helps standardized diagnostic workflows through the reduction of human failure as well as enhanced information-extraction consistency. Taken together, these outcomes establish that LLMs could bridge the gap between raw clinical data to usable diagnostic information to facilitate earlier treatment as well as personalised care.

3.5.2. Risk Assessment and Triage

In addition to screening and detection, LLMs facilitate the risk assessment and triage procedures, especially in mental health applications. By analysing the sentiment, tone, and complexity of language, such models are able to evaluate the psychological distress and guide those with increased risk of suicide or crisis [17]. This feature is particularly useful in online or resource-limited care environments, where fast triage can easily affect clinical outcomes.

The analytical power of the LLMs also helps practitioners in correlating large amounts of patient-reported information to assist with triage decisions regarding who needs immediate intervention as opposed to ongoing monitoring. As the systems evolve, the integration of the triage tools with the telepsychiatry platforms could allow long-term, passive surveillance of mental state with earlier detection of mental health emergencies.

3.5.3. Clinical Reasoning and Decision Support

The LLMs showed increasing abilities to mimic clinical reasoning and facilitate decision-making within psychiatry, cardiology, and oncology. They can facilitate diagnosis and treatment planning in psychiatry, typically producing guideline-concordant reasoning when presenting with clinical vignettes [15,16]. This helps the clinician in challenging diagnostic situations but also provides a useful instrument within medical education, as it enables students to practice structured ways of reasoning.

In cardiology, the LLMs proved effective in accurately answering board-style examination questions and depicting clinical reasoning in case discussions [4]. Such functions advance both

 doi:10.20944/preprints202510.2480.v1

educational use and ongoing professional development with the option to present an adaptive learning environment that resembles clinician-level judgment.

In oncology, the LLMs aid in the synthesis of multidisciplinary cases and tumor board planning with 50–70% agreement with the expert panels [5]. This analytical potential, in turn, leads to the possibility of summarizing variable clinical facts as well as helping with decision-making processes, most specifically where combined therapies need to be incorporated.

3.5.4. Patient Education and Participation

A major advantage of LLMs is their ability to encourage patient engagement by using natural, conversational interfaces. In mental health, conversational agents that use LLMs can provide psychoeducation, empathetic dialogue, and stigma reduction by offering support to individuals who might otherwise avoid seeking for help [2]. These systems can deliver round-the-clock guidance and emotional validation, contributing to improved accessibility and continuity of care. In cardiology, the LLMs were shown to produce accurate and sympathetic responses to patient questions, enhancing compliance with regimens as well as health literacy [4]. Similarly, in oncology, they facilitate the development of patient education materials as well as consent summaries incorporating the most current clinical practice guidelines [5]. Such functionalities endow the patient with reliable, easy-to-read information, enforcing shared decision-making as well as faith in communication with the clinician.

3.5.5. Summary, Documentation, and Research Support

In all areas, LLMs bring significant advantage in automating research and documentation procedures. They help in psychiatry to summarize transcripts of therapy and to produce structured sets of data for Natural Language Processing (NLP) studies, minimizing manual labor and maximizing analytical potential [16]. They help in cardiology and oncology to prepare discharge summaries, clinical notes, and guideline overviews to minimize administrative workload [13,14]. In addition, the models aid clinical researchers in scientific paper writing through the automation of evidence synthesis and paper composition [13]. This aspect can accelerate the publication of clinical evidence, boosting the efficiency of medical scholarship. As the science of LLMs advances, this integration into research workflows could revolutionize the creation, curation, and communication of medical knowledge.

3.6. Challenges of Using LLMs

Diverse medical fields analysed in the selected papers present transversal challenges, needs, or processes where LLMs different techniques can play a meaningful role. Despite their high potential in early detection and symptom classification, LLMs' use and integration in clinical practice requires a number of concerns to be taken into account such as inconsistent accuracy, efficacy, and technical limitations (e.g., hallucinations), bias, model transparency, data privacy, continuous human oversight, ethical concerns and guidelines, data availability, as well as the design and conduction of studies with high methodological quality.

3.6.1. Accuracy, Safety, and Efficacy of LLMs in Real World Settings

Although LLMs showed better results than traditional tools such as machine learning [22,23], and even exhibited capabilities comparable to those of human experts in some cases [24,25], variations in accuracy and output correctness across different tasks persist. As an example, Levkovich and Elyoseph [26], evaluated ChatGPT's performance in assessing suicide risk highlighting that ChatGPT could underestimate or overestimate suicide risks compared to mental health professionals, especially in complex scenarios with high perceived burdensomeness and thwarted belongingness. Advanced models such as GPT-4 have been effective in interpreting clinical and unstructured data to manage, detect, and classify depression [27]. However, studies often rely on fictional clinical

 doi:10.20944/preprints202510.2480.v1

vignettes, limiting the generalisability of these findings in real-world clinical practice [28]. In this context, researchers and industrial stakeholders should gather enough evidence of efficacy and safety—through rigorous clinical studies, testing, and oversight—before deploying LLMs at scale in real-life, in order to prevent causing harm due to unverified performance.

All reviews underscore the enduring risk of hallucinations, wherein models generate information that is incorrect, incomplete, or entirely fabricated. Such outputs—ranging from inaccurate facts to invented citations—pose particular danger in clinical settings, where they may inadvertently mislead healthcare professionals or patients. The reviews consistently identify hallucination control as a fundamental technical challenge that must be addressed prior to any autonomous clinical deployment of LLMs.

3.6.2. Bias and Model Transparency

Training data biases (demographic, geographic, linguistic) may propagate into LLM outputs. Several reviews note age, language and population skews (e.g., social-media datasets over-represent younger, English-speaking users), leading to uneven performance and potential amplification of disparities in care. Detecting and mitigating bias is made harder because training corpora and curation processes for commercial models are often undisclosed. Reviews call for benchmark datasets with diverse, annotated examples and for bias-auditing pipelines. Ensuring demographic inclusivity, representational diversity, and transparency in LLM development is essential to safeguard public trust in AI-driven healthcare systems.

A critical assessment of the safety and trust of LLMs was conducted in included review studies, highlighting the "black box" nature of AI systems. This means that LLMs are associated with limited interpretability and transparency (why the model generated a given answer), which undermines trust and makes clinical validation and regulatory assessment challenging. The reviews recommend model documentation, provenance of training data, auditing, and research into explanatory methods (attention analyses, knowledge graphs, causal embeddings), in order to improve model transparency.

3.6.3. Data Privacy, Security, and Regulatory Challenges

All reviews identify significant challenges related to data protection and regulatory compliance when applying LLMs in healthcare. Cloud-based model architectures risk exposing sensitive health information to third-party servers, while conventional anonymisation is often inadequate, as models may inadvertently reconstruct or reveal personal data. The literature highlights the need for privacy-preserving solutions such as federated learning, encrypted inference, and local on-premise deployments. Regulatory frameworks—including the GDPR and the forthcoming EU AI Act—remain ill-equipped to address GenAI, leaving uncertainties over accountability, liability, and data-control roles. The opacity of commercial models further complicates auditability and certification under existing medical device standards. Reviews therefore call for transparent data-governance mechanisms, regulatory sandboxes, and privacy-by-design principles to ensure both ethical integrity and legal compliance. Ultimately, safeguarding patient privacy is viewed as a prerequisite for the responsible clinical integration of LLMs.

3.6.4. Data Availability and Generalizability

Data scarcity and imbalance have been identified as core barriers to reliable LLM performance in healthcare. Most models are trained or tested on English-language, high-resource datasets—such as PubMed abstracts or online social media—rather than representative clinical data, limiting generalizability across cultures, age groups, and care settings. Under-representation of non-English, minority, and low-income populations leads to systematic performance gaps and potential inequities in care. The opacity of proprietary training corpora further restricts reproducibility and independent bias auditing, raising concerns about compliance with ethical and data protection standards. To address these gaps, the reviews call for open, multilingual benchmark datasets annotated by domain

 doi:10.20944/preprints202510.2480.v1

experts, as well as privacy-preserving data sharing approaches. Strengthening data diversity, transparency, and accessibility is seen as essential to ensure that LLMs achieve equitable, safe, and scientifically valid integration into healthcare practice.

3.6.5. Continuous Human Oversight and Ethical Governance Frameworks

All five reviews converge on the principle that large language models must operate under continuous human supervision when applied in clinical or mental health contexts. While LLMs demonstrate promise in tasks such as information retrieval, triage support, and patient education, their susceptibility to hallucinations, bias, and contextual misinterpretation makes their unsupervised use unsafe. The reviews emphasize that LLMs should function as decision-support tools rather than autonomous agents, complementing but not replacing expert judgment. Ongoing human monitoring is also crucial for detecting subtle model drift or unexpected behavior following software updates. In this direction, structured human-in-the-loop frameworks, with clearly defined escalation procedures for high-risk outputs (e.g., suicide ideation detection, diagnostic recommendations) are required [29].

All five reviews stress that deploying LLMs in healthcare demands clear ethical governance frameworks ensuring transparency, accountability, and respect for patient rights. Current practice often outpaces regulation, leaving uncertainty around responsibility, informed consent, and fairness. Reviews highlight core principles such as accountability (clinicians retain final responsibility), transparency (users must know when AI is involved), non-maleficence (preventing harm through validation), justice (equitable performance across populations), autonomy (informed consent for AI interaction), and data ethics (responsible stewardship and privacy protection). Overall, the literature emphasises that robust, enforceable ethical frameworks—combining technical, institutional, and societal safeguards—are essential to maintain public trust and ensure that GenAI serves healthcare's fundamental moral obligations.

3.6.6. Methodological Quality of LLM Studies

All reviews highlight significant methodological weaknesses in current research on LLMs in healthcare. Most studies rely on retrospective, simulated, or vignette-based data rather than real-world clinical trials, limiting external validity and generalizability. Evaluation protocols vary widely, with inconsistent reporting of datasets, prompts, metrics, and baseline comparisons, making cross-study synthesis difficult. Few investigations employ standardized bias or risk-of-bias tools (e.g., QUADAS-2, PROBAST), and many omit details about model versioning or update cycles—critical for reproducibility. Reviews call for prospective, pre-registered, and multi-institutional studies using transparent methods and benchmark datasets that reflect clinical complexity and population diversity. They also emphasize the need for longitudinal evaluations assessing safety, performance drift, and human–AI interaction over time. Overall, improving methodological rigor and transparency is seen as essential to move the field from proof-of-concept experimentation toward clinically validated, ethically sound, and policy-relevant evidence. A summary of all main outcomes and challenges as highlighted in the included reviews can be seen in Table 2.

Table 2. Summary of outcomes and challenges.

Chronic diseases	Common symptoms/needs	Main outcomes of review studies	Benefits from integration of LLMs into clinical practice	Overall challenges
Mental Health/ Psychiatry	• Depression • Anxiety disorders • Cognitive disorders	• Implementation of LLMs for the detection, diagnosis, classification and management of mental disorders such as depression	• Enhanced efficiency and accessibility • Support in early diagnosis and clinical management	• Concerns about data privacy, ethical implications & bias • Potential discrimination that may exacerbate health disparities

 doi:10.20944/preprints202510.2480.v1

	• Irritability or anger • Hopelessness or helplessness • Loss of interest or pleasure in activities	• Risk assessment and patient triage • Patient support and education • Analysis of clinical data and social media texts	• Support in patient stratification/classification • Exploration of symptomatology or clinical data to assist decision support • Generation of clinically meaningful variables from unstructured data	• Need for regulatory frameworks and transparency, reliability, confidentiality, and trustworthiness in the patient-physician relationship • Need for advanced efficacy, safety & human oversight in LLMs implementation in real world settings • Lack of public datasets for exploring the intersection between the disease & AI • Misleading medical information, medical errors, risk of information misuse or unverified treatments in complex clinical cases
Cardiology	• Chronic heart failure • Hypertension • Cardiac rehabilitation & patient adherence	• Patient education literacy • Clinical reasoning, training • Summarization, documentation • Research writing support	• Correct responses to cardiology questions • Simulation of reasoning for trainee education • Automated synthesis of evidence for clinicians & researchers	• Mimicking clinician reasoning in case discussions • Accurate and empathetic responses to patient questions • Drafting discharge summaries and guideline synopses, reducing documentation burden • Ethical concerns regarding the delivery of medical advice and the authoring of manuscripts • Enabling personalised patient care as well as scale-wide synthesis from the literature
Oncology/Breast Cancer	• Diagnosis t • Survivorship • Psycho-social support	• Oncology/breast cancer management & clinical decision support • Information extraction & evidence retrieval	• Improved data processing time • Standardised diagnostic workflows through the reduction of human errors • Bridging the gap between raw clinical data to usable diagnostic information	• Synthesis of multidisciplinary cases and tumor board planning • Possibility of summarizing variable clinical facts as well as helping with decision-making processes • Summarization use in combined therapies' implementation • Incorporating the most up-to-date clinical practice guidelines • Facilitate earlier treatment as well as personalised care

4. Discussion

4.1. Main Findings

We conducted a review of systematic reviews on the application of LLMs in critical healthcare domains, including cardiology, oncology, and mental health. The main goal was to examine the characteristics, outcomes, and challenges of LLMs, by drawing, exploring, and synthesising the

 doi:10.20944/preprints202510.2480.v1

findings of systematic reviews in this novel area of research. The key finding of the review is that LLMs have emerged as potential enhancers of diagnostic and decision-support processes in cardiology, oncology, and mental health. Nonetheless, the limited rigorous evidence to date underscores the need for robust, real-world research to validate LLM effectiveness and safety in clinical practice.

The predominant research focus was on depression detection and classification, complemented by investigations into suicidal ideation detection, question answering aligned with clinical guidelines, AI-driven mental health interventions such as loneliness support, emotion recognition and classification, and tumor board clinical decision support. More than half of the included studies applied GPT-based models, while BERT and its derivatives were employed in approximately 40% of the studies. The performance of LLMs was variable across their broad spectrum of applications, and therefore no clear evidence of their effectiveness can be demonstrated currently.

Across cardiology, oncology, and mental health, the reviewed evidence indicates that LLMs hold substantial potential to enhance various aspects of clinical care, education, and research. Their strongest demonstrated benefits lie in detection, triage, and decision support, where transformer-based architectures can process vast volumes of unstructured text to identify early markers of disease, assess patient risk, and generate structured diagnostic information with high accuracy.

In mental health, LLMs have been applied to detect depression, anxiety, and suicidality through linguistic and affective cues, enabling scalable, non-intrusive screening beyond traditional clinical environments. In this context, LLMs may facilitate risk stratification and triage, helping clinicians identify individuals in need of early intervention—particularly in resource-limited or telehealth settings [30]. Their capacity for clinical reasoning and decision support has been demonstrated also in cardiology and oncology, where models often produce guideline-concordant responses and may assist in diagnostic synthesis and multidisciplinary planning [31].

Another prominent advantage of LLMs is their ability to enhance patient education and engagement through empathetic, conversational interfaces that provide accessible health information and encourage adherence and shared decision-making [32]. Additionally, LLMs streamline documentation, summarization, and research workflows, reducing administrative burden and accelerating knowledge generation. By automating report generation, literature synthesis, and evidence summarisation, they can increase efficiency across clinical research domains [33].

Despite the potential of LLMs in improving clinical care, several limitations and challenges toward their wide adoption can be highlighted. Model output correctness and hallucinations remain fundamental challenges in applying LLMs to healthcare. Models frequently produce plausible but inaccurate or fabricated information, including erroneous clinical facts or nonexistent references [34]. Such hallucinations undermine reliability and pose safety risks when used for diagnosis or patient communication. The literature attributes these errors to limitations in the inherent logical structure of LLMs, training data, probabilistic text generation, and lack of factual grounding.

Furthermore, the included reviews consistently highlighted bias and lack of transparency as critical obstacles to the safe deployment of LLMs in healthcare. In several cases, the opacity of proprietary model architectures and datasets prevents independent auditing, bias detection, and accountability [35]. In this direction, transparent reporting of data provenance, inclusive and diverse training corpora, and bias-auditing frameworks that systematically evaluate fairness and performance across demographic and linguistic groups before clinical implementation, are important steps toward the improvement of LLMs reliability.

The reviewed evidence underscores major challenges related to data privacy, availability, human oversight, and methodological rigor in LLM research. Most models rely on proprietary or web-scraped data with uncertain consent and provenance, raising privacy and compliance concerns under frameworks such as GDPR. Simultaneously, the lack of open, diverse, expert-labelled, and representative datasets limits reproducibility and generalisability across languages and populations. The reviews stress that continuous human oversight is indispensable—LLMs should function as decision-support tools, with clinicians validating outputs and managing escalation for high-risk

 doi:10.20944/preprints202510.2480.v1

content. Methodologically, the current literature remains exploratory, dominated by simulated or retrospective designs with inconsistent evaluation metrics and scarce real-world validation. To advance clinical integration responsibly, the field must prioritize transparent data governance, ethical data sharing, standardized evaluation protocols, and prospective trials that rigorously test performance, safety, and reliability in real-world healthcare settings.

4.2. Limitations

This review should be interpreted within the context of its limitations. The authors used a limited set of terms for the search of the literature. These were related to LLMs, chatbots, and conversational agents. Another limitation is that our search was conducted in a limited number of bibliographic databases (PubMed and Scopus). This might have resulted in the omission of other relevant studies. The review was based on the findings from systematic reviews of LLM studies and different populations in terms of disease, age, education and socioeconomic level were studied in different settings, which prevented the conduction of a meta-analysis. The limitations of the included systematic reviews (introduced for example in their inclusion and exclusion criteria) might have affected the representation of the progress of LLMs. This review presented features and outcomes of LLMs for cardiovascular disease, cancer, and mental disorders, considering their global burden. LLMs for other chronic conditions such as diabetes, arthritis, or liver disease were not examined. Our literature search was restricted to reviews published after 2022, due to the landmark emergence of ChatGPT. Different results could emerge if older studies were also included. The generalisability of the findings is restricted by the fact that only a small number of studies was found to be eligible for inclusion in this review.

5. Conclusion

In conclusion, the collective evidence from recent systematic reviews demonstrates that LLMs hold substantial promise for enhancing healthcare through improved detection, triage, decision support, and patient engagement. Overall, the reviews converge on the view that while LLMs are not yet ready for autonomous deployment in clinical settings, their augmentation of human expertise—from early detection and clinical reasoning to patient communication and research support—positions them as valuable tools for improving efficiency, consistency, and accessibility of healthcare services. Yet, their clinical integration remains constrained by persistent challenges, including hallucinations, bias, lack of transparency, data privacy risks, and limited methodological robustness. Current research is largely exploratory, emphasising the need for rigorous, real-world evaluations supported by transparent data governance and ethical oversight. Ensuring reliability, fairness, and human supervision will be essential to build trust and safeguard patient welfare. As these models evolve, interdisciplinary collaboration among clinicians, data scientists, ethicists, and policymakers will be crucial to translate technical innovation into safe, equitable, and accountable clinical practice. Properly governed and validated, LLMs could become transformative instruments in shaping the future of evidence-based, human-centred healthcare.

Author Contributions: Conceptualization, A.T.; methodology, A.T., S.S., S.K., An.A., E.L.; writing—original draft preparation, A.T.; writing—review and editing, all authors; supervision, A.T.; project administration, A.T.; funding acquisition, A.T., G.M., At.A., K.V., D.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the European Union's Horizon 2020 research and innovation program under grant agreement No 101080430 (AI4HF), and the European Union's Horizon 2020 research and innovation program under grant agreement No 101137301 (COMFORTage).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

 doi:10.20944/preprints202510.2480.v1

Data Availability Statement: No new data were created or analyzed in this study.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Appendix A

Table A1. Characteristics of individual studies included in the reviews.

Study	Target Disease	Application	LLM Used	Data Sources	Accuracy	Key Outcomes
Kusunose et al. [36]	CVD	Question-answering based on Japanese Hypertension guidelines	GPT 3.5	Japanese Society of Hypertension Guidelines for the Management of Hypertension	64.50%	ChatGPT performed well on clinical questions, performance on hypertension treatment guidelines topics was less satisfactory
Rizwan et al. [37]	CVD	Question-answering on treatment and management plans	GPT 4.0	Hypothetical questions simulating clinical consultation	N/A	Out of the 10 clinical scenarios inserted in ChatGPT, eight were perfectly diagnosed
Skalidis et al. [38]	CVD	Ability in European Cardiology exams	GPT (Version N/A)	Exam questions from the ESC website, StudyPRN and Braunwald's Heart Disease Review and Assessment	58.80%	Results demonstrate that ChatGPT succeeds in the European Cardiology exams
Van Bulck et al. [39]	CVD	Question-answering on common cardiovascular diseases	GPT (Version N/A)	Virtual patient questions	N/A	Experts considered ChatGPT-generated responses trustworthy and valuable, with few considering them dangerous. Forty percent of the experts found ChatGPT responses more valuable than Google
Williams et al. [40]	CVD	Question-answering on cardiovascular computed tomography	GPT 3.5	Questions from the Society of Cardiovascular Computed Tomography 2023 programme as well as questions about high risk plaque (HRP), quantitative plaque analysis, and how AI will	N/A	The answers to debate questions were plausible and provided reasonable summaries of the key debate points

 doi:10.20944/preprints202510.2480.v1

				transform cardiovascular CT		
Choi et al. [41]	Breast Cancer	Evaluation of the time and cost of developing prompts using LLMs, tailored to extract clinical factors in breast cancer patients	GPT 3.5	Data from reports of surgical pathology and ultrasound from 2931 breast cancer patients who underwent radiotherapy from 2020 to 2022	87.70%	Developing and processing the prompts took 3.5 hours and 15 minutes, respectively. Utilizing the ChatGPT application programming interface cost US $65.8 and when factoring in the estimated wage, the total cost was US $95.4
Griewing et al. [42]	Breast Cancer	Concordance to tumor board clinical decisions	GPT 3.5	Fictious patient data with clinical and diagnostic data	50%-95%	ChatGPT 3.5 can provide treatment recommendations for breast cancer patients that are consistent with multidisciplinary tumor board decision making of a gynecologic oncology center in Germany
Haver et al. [43]	Breast Cancer	Question-answering on breast cancer prevention and screening	GPT 3.5	25 questions addressing fundamental concepts related to breast cancer prevention and screening	88%	ChatGPT provided appropriate responses for most questions posed about breast cancer prevention and screening, as assessed by fellowship-trained breast radiologists
Lukac et al. [44]	Breast Cancer	Tumor board clinical decision support	GPT 3.5	Tumor characteristics and age of the 10 consecutive pretreatment patient cases	64.20%	ChatGPT provided mostly general answers based on inputs, generally in agreement with the decision of MDT
Rao et al. [45]	Breast Cancer	Question-answering based on American Collegy of Radilology recommendations	GPT 4.0, GPT 3.5	Prompting mechanisms and clinical presentations based on American College of Radiology	88.9%-98.4%	ChatGPT displays impressive accuracy in identifying appropriateness of common imaging modalities for breast cancer screening and breast pain
Sorin et al. [46]	Breast Cancer	Tumor board clinical decision support	GPT 3.5	Clinical and diagnostic data of 10 patients	70%	Chatbot's clinical recommendations were in-line with those of the tumor board in 70% of cases
Abilkaiyrkyzy et al. [47]	Mental Disease	Mental illness detection	BERT	219 E-DAIC participants	69%	Chatbot effectively detects and classifies mental health issues, highly usable for

 doi:10.20944/preprints202510.2480.v1

		using a chatbot				reducing barriers to mental health care
Adarsh et al. [48]	Mental Disease	Depression sign detection using BERT	BERT-small	Social media texts	63.60%	Enhanced BERT model accurately classifies depression severity from social media texts, understanding nuances better than others
Alessa and Al-Khalifa [49]	Mental Disease	Mental health interventions using CAs for the elderly supported by LLMs	ChatGPT; Google Cloud API	Record of interactions with CA; results of the human experts' assessment	N/A	The proposed ChatGPT-based system effectively serves as a companion for elderly individuals, helping to alleviate loneliness and social isolation. Preliminary evaluations showed that the system could generate relevant responses tailored to elderly personas
Beredo and Ong [50]	Mental Disease	Mental health interventions using CAs supported by LLMs	EREN;MHBot;PERMA	Empatheticdialogues (24,850 conversations); Well-Being Conversations;Perma Lexica	N/A	This study successfully demonstrated a hybrid conversation model, which combines generative and retrieval approaches to improve language fluency and empathetic response generation in chatbots. This model, tested through both automated metrics and human evaluation, showed that the medium variation of the FTER model outperformed the vanilla DialoGPT in perplexity and that the human-likeness, relevance, and empathetic qualities of responses were significantly enhanced, making VHope a more competent CA with empathetic abilities
Berrezueta-Guzman et al. [51]	Mental Disease	Evaluation of the efficacy of ChatGPT in mental intensive treatment	ChatGPT	Evaluations from 10 attention deficit hyperactivity disorder (ADHD) therapy experts and interactions between therapists and the custom ChatGPT	N/A	This paper found that the custom ChatGPT demonstrated strong capabilities in engaging language use, maintaining interest, promoting active participation, and fostering a positive atmosphere in ADHD therapy sessions, with high ratings in communication and language. However, areas needing improvement were

 doi:10.20944/preprints202510.2480.v1

						identified, particularly in confidentiality and privacy, cultural and sensory sensitivity, and handling nonverbal cues
Blease et al. [52]	Mental Disease	Evaluation of psychiatrists' perceptions of the LLMs	ChatGPT; Bard;Bing AI	Survey responses from 138 APA members on LLM chatbot use in psychiatry	N/A	This paper found that over half of psychiatrists used AI tools like ChatGPT for clinical questions, with nearly 70% agreeing on improved documentation efficiency and almost 90% indicating a need for more training while expressing mixed opinions on patient care impacts and privacy concerns
Bokolo et al. [23]	Mental Disease	Depression detection from Twitter	RoBERTa, DeBERTa	632,000 tweets	97.48%	Transformer models like RoBERTa excel in depression detection from Twitter data, outperforming traditional ML approaches
Crasto et al. [53]	Mental Disease	Mental health interventions using CAs supported by LLMs	DialoGPT	Counselchat (includes tags of illness);question answers from 100 college students	N/A	The DialoGPT model, demonstrating higher perplexity and preferred by 63% of college participants for its human-like and empathetic responses, was chosen as the most suitable system for addressing student mental health issues
Dai et al. [18]	Mental Disease	Psychiatric patient screening	BERT, DistilBERT, ALBERT, ROBERTa	500 EHRs	Accuracy not reported, F1 0.830	BERT models, especially with feature dependency, effectively classify psychiatric conditions from EHRs
Danner et al. [54]	Mental Disease	Detecting depression using LLMs through clinical interviews	BERT;GPT-3.5;ChatGPT-4	DAIC-WOZ;Extended-DAIC;simulated data	78%	The study assessed the abilities of GPT-3.5-turbo and ChatGPT-4 on the DAIC-WOZ dataset, which yielded F1 scores of 0.78 and 0.61 respectively, and a custom BERT model, extended-trained on a larger dataset, which achieved an F1 score of 0.82 on the Extended-DAIC dataset, in recognizing depression in text
Dergaa et al. [55]	Mental Disease	Simulated mental health assessments and	GPT-3.5	Fictional patient data; 3 scenarios	N/A	ChatGPT showed limitations in complex medical scenarios, underlining its unpreparedness for

 doi:10.20944/preprints202510.2480.v1

		interventions with ChatGPT				standalone use in mental health practice
Diniz et al. [56]	Mental Disease	Detecting suicidal ideation using LLMs through Twitter texts	BERT model for Portuguese;Multilingual BERT (base);BERTimbau	Non-clinical texts from tweets (user posts of the online social network Twitter)	95%	The Boamente system demonstrated effective text analysis for suicidal ideation with high privacy standards and actionable insights for mental health professionals. The best-performing BERTimbau Large model (accuracy: 0.955; precision: 0.961; F-score: 0.954; AUC: 0.954) significantly excelled in detecting suicidal tendencies, showcasing robust accuracy and recall in model evaluations
D'Souza et al. [21]	Mental Disease	Responding to psychiatric case vignettes with diagnostic and management strategies	GPT-3.5	Fictional patient data from clinical case vignettes; 100 cases	61%	ChatGPT 3.5 showed high competence in handling psychiatric case vignettes, with strong diagnostic and management strategy generation
Elyoseph et al. [57]	Mental Disease	Evaluating emotional awareness compared to general population norms	GPT-3.5	Fictional scenarios from the LEAS; 750 participants	85%	ChatGPT showed higher emotional awareness compared to the general population and improved over time
Elyoseph et al. [58]	Mental Disease	Assessing suicide risk in fictional scenarios and comparing to professional evaluations	GPT-3.5	Fictional patient data; text vignettescompared to 379 professionals	N/A	ChatGPT underestimated suicide risks compared to mental health professionals, indicating the need for human judgment in complex assessments
Elyoseph et al. [24]	Mental Disease	Evaluating prognosis in depression compared to other LLMs and professionals	GPT-3.5, GPT-4	Fictional patient data; text vignettescompared to 379 professionals	N/A	ChatGPT-3.5 showed a more pessimistic prognosis in depression compared to other LLMs and mental health professionals

 doi:10.20944/preprints202510.2480.v1

Esackimuthu et al. [59]	Mental Disease	Depression detection from social media text	ALBERT base v1	ALBERT base v1 data	50%	ALBERT shows potential in detecting depression signs from social media texts but faces challenges due to complex human emotions
Farhat et al. [60]	Mental Disease	Evaluation of ChatGPT as a complementary mental health resource	ChatGPT	Responses generated by ChatGPT	N/A	ChatGPT displayed significant inconsistencies and low reliability when providing mental health support for anxiety and depression, underlining the necessity of validation by medical professionals and cautious use in mental health contexts
Farruque et al. [61]	Mental Disease	Depression level detection modelling	Mental BERT (MBERT)	13,387 Reddit samples	Accuracy not reported, F1 0.81	MBERT enhanced with text excerpts significantly improves depression level classification from social media posts
Farruque et al. [62]	Mental Disease	Depression symptoms modelling from Twitter	BERT, Mental-BERT	6077 tweets and 1500 annotated tweets	Accuracy not reported, F1 0.45	Semi-supervised learning models, iteratively refined with Twitter data, improve depression symptom detection accuracy
Friedman and Ballentine [63]	Mental Disease	Evaluation of LLMs in data-driven discovery: correlating sentiment changes with psychoactive experiences	BERTowid;BERTiment	Erowid testimonials;drug receptor affinities;brain gene expression data;58K annotated Reddit posts	N/A	This paper found that LLM methods can create unified and robust quantifications of subjective experiences across various psychoactive substances and timescales. The representations learned are evocative and mutually confirmatory, indicating significant potential for LLMs in characterizing psychoactivity
Ghanadian et al. [64]	Mental Disease	Suicidal ideation detection using LLMs through social media texts	ALBERT; DistilBERT; ChatGPT; Flan-T5;Llama	UMD Dataset; Synthetic Datasets (Generated using LLMs like Flan-T5 and Llama2, these datasets augment the UMD dataset to enhance model performance)	87%	The synthetic data-driven method achieved consistent F1-scores of 0.82, comparable to real-world data models yielding F1-scores between 0.75 and 0.87. When 30% of the real-world UMD dataset was combined with the synthetic data, the performance significantly improved, reaching an F1-score of 0.88 on the UMD test set. This result highlights the effectiveness of synthetic data in addressing data scarcity

 doi:10.20944/preprints202510.2480.v1

						and enhancing model performance
Hadar-Shoval et al. [65]	Mental Disease	Differentiating emotional responses in BPD and SPD scenarios using mentalising abilities	GPT-3.5	Fictional patient data (BPD and SPD scenarios); AI-generated data	N/A	ChatGPT effectively differentiated emotional responses in BPD and SPD scenarios, showing tailored mentalizing abilities
Hayati et al. [66]	Mental Disease	Detecting depression by Malay dialect speech using LLMs	GPT-3	Interviews with 53 adults fluent in Kuala Lumpur (KL), Pahang, or Terengganu Malay dialects	73%	GPT-3 was tested on three different dialectal Malay datasets (combined, KL, and non-KL). It performed best on the KL dataset with a max_example value of 10, which achieved the highest overall performance. Despite the promising results, the non-KL dataset showed the lowest performance, suggesting that larger or more homogeneous datasets might be necessary for improved accuracy in depression detection tasks
He et al. [67]	Mental Disease	Evaluation of CAs handling counseling for people with autism supported by LLMs	ChatGPT	Public available data from the web-based medical consultation platform DXY	N/A	The study found that 46.86% of assessors preferred responses from physicians, 34.87% favored ChatGPT, and 18.27% favored ERNIE Bot. Physicians and ChatGPT showed higher accuracy and usefulness compared to ERNIE Bot, while ChatGPT outperformed both in empathy. The study concluded that while physicians' responses were generally superior, LLMs like ChatGPT can provide valuable guidance and greater empathy, though further optimization and research are needed for clinical integration
Hegde et al. [68]	Mental Disease	Depression detection using supervised learning	Ensemble of ML classifiers, BERT	Social media text data	Accuracy not reported, F1 0.479	BERT-based Transfer Learning model outperforms traditional ML classifiers in detecting depression from social media texts

 doi:10.20944/preprints202510.2480.v1

Heston T.F. et al. [69]	Mental Disease	Simulating depression scenarios and evaluating AI's responses	GPT-3.5	Fictional patient data; 25 conversational agents	N/A	ChatGPT-3.5 conversational agents recommended human support at critical points, highlighting the need for AI safety in mental health
Hond et al. [70]	Mental Disease	Early depression risk detection in cancer patients	BERT	16,159 cancer patients' EHR data	Accuracy not reported, AUROC 0.74	Machine learning models predict depression risk in cancer patients using EHRs, with structured data models performing best
Howard et al. [71]	Mental Disease	Detecting suicidal ideation using LLMs through social media texts	DeepMoji;Universal Sentence Encoder; GPT-1	1588 labeled posts from the Computational Linguistics and Clinical Psychology 2017 shared task	Accuracy not reported, F1 0.414	The top-performing system, utilizing features derived from the GPT-1 model fine-tuned on over 150,000 unlabeled Reachout.com posts, achieved a new state-of-the-art macro-averaged F1 score of 0.572 on the CLPsych 2017 task without relying on metadata or preceding posts. However, error analysis indicated that this system frequently misses expressions of hopelessness
Hwang et al. [72]	Mental Disease	Generating psychodynamic formulations in psychiatry based on patient history	GPT-4	Fictional patient data from published psychoanalytic literature; 1 detailed case	N/A	GPT-4 successfully created relevant and accurate psychodynamic formulations based on patient history
Ilias et al. [73]	Mental Disease	Stress and depression identification in social media	BERT, MentalBERT	Public datasets	Accuracy not reported, F1 0.73	Extra-linguistic features improve calibration and performance of models in detecting stress and depression from texts
Janatdoust et al. [74]	Mental Disease	Depression signs detection from social media text	Ensemble of BERT, ALBERT, DistilBERT, RoBERTa	16,632 social media comments	61%	Ensemble models effectively classify depression signs from social media, utilizing multiple language models for improved accuracy
Kabir et al. [75]	Mental Disease	Depression severity detection from tweets	BERT, DistilBERT	40,191 tweets	Accuracy not reported, AURO	Models effectively classify social media texts into depression severity categories, with high confidence and accuracy

 doi:10.20944/preprints202510.2480.v1

					C 0.74-0.86	
Kumar et al. [76]	Mental Disease	Evaluation of GPT 3 in mental health intervention	GPT 3	209 participants responses, with 189 valid responses after filtering	N/A	This paper found that interaction with either of the chatbots improved participants' intent to practice mindfulness again, while the tutorial video enhanced their overall experience of the exercise. These findings highlighted the potential promise and outlined directions for exploring the use of LLM-based chatbots for awareness-related interventions
Lam et al. [77]	Mental Disease	Multi-modal depression detection	Transformer, 1D CNN	189 DAIC-WOZ participants	Accuracy not reported, F1 0.87	Multi-modal models combining text and audio data effectively detect depression, enhanced by data augmentation
Lau et al. [22]	Mental Disease	Depression severity assessment	Prefix-tuned LLM	189 clinical interview transcripts	Accuracy not reported, RMSE 4.67	LLMs with prefix-tuning significantly enhance depression severity assessment, surpassing traditional methods
Levkovich and Elyoseph [25]	Mental Disease	Diagnosing and treating depression, comparing GPT models with primary care physicians	GPT-3.5, GPT-4	Fictional patient data from clinical case vignettes; repeated multiple times for consistency	N/A	ChatGPT aligned with guidelines for depression management, contrasting with primary care physicians and showing no gender or socioeconomic biases
Levkovich and Elyoseph [26]	Mental Disease	Evaluating suicide risk assessments by GPT models and mental health professionals	GPT-3.5, GPT-4	Fictional patient data; text vignettes compared to 379 professionals	N/A	GPT-4's evaluations of suicide risk were similar to mental health professionals, though with some overestimations and underestimations
Li et al. [78]	Mental Disease	Evaluating performance on psychiatric licensing exams and diagnostics	GPT-4, Bard and Llama-2	Fictional patient data in exam and clinical scenario questions; 24 experienced psychiatrists	69%	GPT-4 outperformed other models in psychiatric diagnostics, closely matching the capabilities of human psychiatrists

 doi:10.20944/preprints202510.2480.v1

Liyanage et al. [79]	Mental Disease	Data augmentation for wellness dimension classification in Reddit posts	GPT-3.5	Real patient data from Reddit posts; 3,092 instances, post-augmentation 4376 records	69%	ChatGPT models effectively augmented Reddit post data, significantly improving classification performance for wellness dimensions
Lossio-Ventura et al. [20]	Mental Disease	Evaluations of LLMs for sentiment analysis through social media texts	ChatGPT; Open Pre-Trained Transformers (OPT)	NIH Data Set;Stanford Data Set	86%	This paper revealed high variability and disagreement among sentiment analysis tools when applied to health-related survey data. OPT and ChatGPT demonstrated superior performance, outperforming all other tools. Moreover, ChatGPT outperformed OPT, achieving a 6% higher accuracy and a 4% to 7% higher F-measure
Lu et al. [80]	Mental Disease	Depression detection via conversation turn classification	BERT, transformer encoder	DAIC dataset	Accuracy not reported, F1 0.75	Novel deep learning framework enhances depression detection from psychiatric interview data, improving interpretability
Ma et al. [81]	Mental Disease	Evaluation of mental health intervention CAs supported by LLMs	GPT-3	120 Reddit posts (2913 user comments)	N/A	The study highlighted that CAs like Replika, powered by LLMs, offered crucial mental health support by providing immediate, unbiased assistance and fostering self-discovery. However, they struggled with content filtering, consistency, user dependency, and social stigma, underscoring the importance of cautious use and improvement in mental wellness applications
Mazumdar et al. [82]	Mental Disease	Classifying mental health disorders and generating explanations	GPT-3, BERT-large,MentalBERT, ClinicBERT, and PsychBERT	Real patient data sourced from Reddit posts	87%	GPT-3 outperformed other models in classifying mental health disorders and generating explanations, showing promise for AI-IoMT deployment
Metzler et al. [83]	Mental Disease	Detecting suicidal ideation	BERT;XLNet	3202 English tweets	88.50%	BERT achieved F1-scores of 0.93 for accurately labeling tweets as about suicide and

 doi:10.20944/preprints202510.2480.v1

		using LLMs through Twitter texts				0.74 for off-topic tweets in the binary classification task. Its performance was similar to or exceeded human performance and matched that of state-of-the-art models on similar tasks
Owen et al. [84]	Mental Disease	Depression signal detection in Reddit posts	BERT, MentalBERT	Reddit datasets	Accuracy not reported, F1 0.64	Effective identification of depressive signals in online forums, with potential for early intervention
Parker et al. [85]	Mental Disease	Providing information on bipolar disorder and generating creative content	GPT-3	N/A	N/A	GPT-3 provided basic material on bipolar disorders and creative song generation, but lacked depth for scientific review
Perlis et al. [28]	Mental Disease	Evaluation of GPT-4 for clinical decision support in bipolar depression	GPT-4 turbo (gpt-4-1106-preview)	Recommendations generated by the augmented GPT-4 model and responses from clinicians treating bipolar disorder	50.80%	This paper found that the augmented GPT-4 model had a Cohen's kappa of 0.31 with expert consensus, identifying the optimal treatment in 50.8% of cases and placing it in the top 3 in 84.4% of cases. In contrast, the base model had a Cohen's kappa of 0.09 and identified the optimal treatment in 23.4% of cases, highlighting the enhanced performance of the augmented model in aligning with expert recommendations
Poświata and Perełkiewicz [86]	Mental Disease	Depression sign detection using RoBERTa	RoBERTa, DepRoBERTa	RoBERTa models' data	Accuracy not reported, F1 0.583	RoBERTa and DepRoBERTa ensemble excels in classifying depression signs, securing top performance in a competitive environment
Pourkeyvan et al. [19]	Mental Disease	Mental health disorder prediction from Twitter	BERT models from Hugging Face	11,890,632 tweets and 553 bio-descriptions	97%	Superior detection of depression symptoms from social media, demonstrating the efficacy of advanced NLP models
Sadeghi et al. [87]	Mental Disease	Detecting depression using LLMs through interviews	GPT-3.5-Turbo;RoBERTa	E-DAIC (219 participants)	N/A	The study achieved its lowest error rates, a Mean Absolute Error (MAE) of 3.65 on the dev set and 4.26 on the test set, by fine-tuning DepRoBERTa with a specific

 doi:10.20944/preprints202510.2480.v1

						prompt, outperforming manual methods and highlighting the potential of automated text analysis for depression detection
Schubert et al. [88]	Mental Disease	Evaluation of LLMs' performance on neurology board-style examinations	ChatGPT 3.5;ChatGPT 4.0	A question bank from an educational company with 2036 questions that resemble neurology board questions	85%	ChatGPT 4.0 excelled over ChatGPT 3.5, achieving 85.0% accuracy versus 66.8%. It surpassed human performance in specific areas and exhibited high confidence in responses. Longer questions tended to result in more incorrect answers for both models
Senn et al. [89]	Mental Disease	Depression classification from interviews	BERT, RoBERTa, DistilBERT	189 clinical interviews	Accuracy not reported, F1 0.93	Ensembles of BERT models enhance depression detection robustness in clinical interviews
Singh and Motlicek [90]	Mental Disease	Depression level classification using BERT, RoBERTa, XLNet	Ensemble of BERT, RoBERTa, XLNet	Ensemble of models	54%	Ensemble model accurately classifies depression levels from social media text, ranking highly in competitive settings
Sivamanikandan S. et al. [91]	Mental Disease	Depression level classification	DistilBERT, RoBERTa, ALBERT	Social media posts	Accuracy not reported, F1 0.457	Transformer models classify depression levels effectively, with RoBERTa achieving the best performance
Spallek et al. [92]	Mental Disease	Providing educational material on mental health and substance use	GPT-4	Real-world queries from mental health and substance use portals; 10 queries	N/A	GPT-4's outputs were substandard compared to expert materials in terms of depth and adherence to communication guidelines
Stigall et al. [93]	Mental Disease	Emotion classification using LLMs through social media texts	EmoBERTTiny	A collection of publicly available datasets hosted on Kaggle and Huggingface	93.14% (sentiment analysis), 85.46% (emotion analysis)	EmoBERTTiny outperformed pre-trained and state-of-the-art models in all metrics and computational efficiency, achieving 93.14% accuracy in sentiment analysis and 85.48% in emotion classification. It processes a 256-token context window in 8.04ms post-tokenization and 154.23ms total processing speed
Suri et al. [94]	Mental Disease	Depressive tendencies	BERT	5997 tweets	97%	Multimodal BERT frameworks significantly

 doi:10.20944/preprints202510.2480.v1

		detection using multimodal data				enhance detection of depressive tendencies from complex social media data
Tao et al. [95]	Mental Disease	Detecting anxiety and depression using LLMs through dialogs in real-life scenarios	ChatGPT	Speech data from nine Q&A tasks related to daily activities (75 patients with anxiety and 64 patients with depression)	67.62%	This paper introduced a virtual interaction framework using LLMs to mitigate negative psychological states. Analysis of Q&A dialogues demonstrated ChatGPT's potential in identifying depression and anxiety. To enhance classification, four language features, including prosodic and speech rate, positively impacted classification
Tey et al. [96]	Mental Disease	Pre- and post-depressive detection from tweets	BERT, supplemented with emoji decoding	Over 3.5 million tweets	Accuracy not reported, F1 0.90	Augmented BERT model classifies Twitter users into depressive categories, enhancing early depression detection
Toto et al. [97]	Mental Disease	Depression screening using audio and text	AudiBERT	189 clinical interviews	Accuracy not reported, F1 0.92	AudiBERT outperforms traditional and hybrid models in depression screening, utilizing multimodal data
Vajre et al. [98]	Mental Disease	Detecting mental health using LLMs through social media texts	PsychBERT	Twitter hashtags and Subreddit (6 domains: anxiety, mental health, suicide, etc)	Accuracy not reported, F1 0.63	The study identified PsychBERT as the highest-performing model, achieving an F1 score of 0.98 in a binary classification task and 0.63 in a more challenging multi-class classification task, indicating its superiority in handling complex mental health-related data. Additionally, PsychBERT's explainability was enhanced by using the Captum library, which confirmed its ability to accurately identify key phrases indicative of mental health issues
Verma et al. [99]	Mental Disease	Detecting depression using LLMs through textual data	RoBERTa	Mental health corpus;	96.86%	The study successfully used a RoBERTa-base model to detect depression with a high accuracy of 96.86%, showcasing the potential of AI in identifying mental health issues through linguistic analysis

 doi:10.20944/preprints202510.2480.v1

Wan et al. [100]	Mental Disease	Family history identification in mood disorders	BERT–CNN	12,006 admission notes	97%	High accuracy in identifying family psychiatric history from EHRs, suggesting utility in understanding mood disorders
Wang et al. [101]	Mental Disease	Enhancing depression diagnosis and treatment through the use of LLMs	LLaMA-7B;ChatGLM-6B;Alpaca;LLMs+ Knowledge	Chinese Incremental Pre-training Dataset	N/A	The study assessed LLMs' performance in mental health, emphasizing safety, usability, and fluency and integrating mental health knowledge to improve model effectiveness, enabling more tailored dialogues for treatment while ensuring safety and usability
Wang et al. [27]	Mental Disease	Detecting depression using LLMs through microblogs	BERT;RoBERTa;XLNet	13,993 microblogs collected from the Sina Weibo	Accuracy not reported, F1 0.424	RoBERTa achieved the highest macro-averaged F1 score of 0.424 for depression classification, while BERT scored the highest micro-averaged F1 score of 0.856. Pretraining on an in-domain corpus improved model performance
Wei et al. [102]	Mental Disease	Evaluation of ChatGPT in psychiatry	ChatGPT	Theoretical analysis and literature reviews	N/A	The paper found ChatGPT useful in psychiatry, stressing ethical use and human oversight, while noting challenges in accuracy and bias, positioning AI as a supportive tool in care
Wu et al. [103]	Mental Disease	Expanding dataset of Post-Traumatic Stress Disorder (PTSD) using LLMs	GPT- 3.5 Turbo	E-DAIC (219 participants)	Accuracy not reported, F1 0.63	This paper demonstrated that two novel text augmentation frameworks using LLMs significantly improved PTSD diagnosis by addressing data imbalances in NLP tasks. The zero-shot approach, which generated new standardized transcripts, achieved the highest performance improvements, while the few-shot approach, which rephrased existing training samples, also surpassed the original dataset's efficacy
Yongsatianchot et al. [104]	Mental Disease	Evaluation of LLMs' perception of emotion	Text-davinci-003;ChatGPT;GPT-4	Responses from three OpenAI LLMs to the Stress and Coping Process Questionnaire	N/A	The study applied the SCPQ to three OpenAI LLMs—davinci-003, ChatGPT, and GPT-4—and found that while their responses aligned with human dynamics of appraisal and coping, they did not vary

 doi:10.20944/preprints202510.2480.v1

					across key appraisal dimensions as predicted and differed significantly in response magnitude. Notably, all models reacted more negatively than humans to negative scenarios, potentially influenced by their training processes
Zhang et al. [105]	Mental Disease	Detecting depression trends using LLMs through Twitter texts	RoBERTa ;XLNet	2575 Twitter users with depression identified via tweets and profiles	78.90%	This study developed a fusion model that accurately classified depression among Twitter users with 78.9% accuracy. It identified key linguistic and behavioral indicators of depression and demonstrated that depressive users responded to the pandemic later than controls. The findings suggest the model's effectiveness in noninvasively monitoring mental health trends during major events like COVID-19

References

1. V. Carchiolo, M. Malgeri, Trends, Challenges, and Applications of Large Language Models in Healthcare: A Bibliometric and Scoping Review, Futur. Internet 2025, Vol. 17, Page 76. 17 (2025) 76. doi:10.3390/FI17020076.
2. E. Kasneci, K. Sessler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser, G. Groh, S. Günnemann, E. Hüllermeier, S. Krusche, G. Kutyniok, T. Michaeli, C. Nerdel, J. Pfeffer, O. Poquet, M. Sailer, A. Schmidt, T. Seidel, M. Stadler, J. Weller, J. Kuhn, G. Kasneci, ChatGPT for good? On opportunities and challenges of large language models for education, Learn. Individ. Differ. 103 (2023) 102274. doi:10.1016/J.LINDIF.2023.102274.
3. S. Volkmer, A. Meyer-Lindenberg, E. Schwarz, Large language models in psychiatry: Opportunities and challenges, Psychiatry Res. 339 (2024) 116026. doi:10.1016/J.PSYCHRES.2024.116026.
4. G. Quer, E.J. Topol, The potential for large language models to transform cardiovascular medicine, Lancet Digit. Heal. 6 (2024) e767 – e771. doi:10.1016/S2589-7500(24)00151-1.
5. G.M. Iannantuono, D. Bracken-Clarke, C.S. Floudas, M. Roselli, J.L. Gulley, F. Karzai, Applications of large language models in cancer care: current evidence and future perspectives, Front. Oncol. 13 (2023) 1268915. doi:10.3389/FONC.2023.1268915.
6. S. Shool, S. Adimi, R. Saboori Amleshi, E. Bitaraf, R. Golpira, M. Tara, A systematic review of large language model (LLM) evaluations in clinical medicine, BMC Med. Inform. Decis. Mak. 25 (2025) 1–11. doi:10.1186/S12911-025-02954-4/FIGURES/2.
7. W. Hussain, G. Khoriba, S. Maity, M. Jyoti Saikia, Large Language Models in Healthcare and Medical Applications: A Review, Bioeng. 2025, Vol. 12, Page 631. 12 (2025) 631. doi:10.3390/BIOENGINEERING12060631.
8. S. Bedi, Y. Liu, L. Orr-Ewing, D. Dash, S. Koyejo, A. Callahan, J.A. Fries, M. Wornow, A. Swaminathan, L.S. Lehmann, H.J. Hong, M. Kashyap, A.R. Chaurasia, N.R. Shah, K. Singh, T. Tazbaz, A. Milstein, M.A. Pfeffer, N.H. Shah, Testing and Evaluation of Health Care Applications of Large Language Models: A Systematic Review., JAMA. (2024). doi:10.1001/jama.2024.21700.

 doi:10.20944/preprints202510.2480.v1

9. D. Moher, A. Liberati, J. Tetzlaff, D.G. Altman, T.P. Group, Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement, PLOS Med. 6 (2009) 1–6. doi:10.1371/journal.pmed.1000097.
10. J.P. Higgins, D.G. Altman, Assessing Risk of Bias in Included Studies, in: Cochrane Handb. Syst. Rev. Interv., John Wiley & Sons, Ltd., Chichester, UK, n.d.: pp. 187–241. doi:10.1002/9780470712184.ch8.
11. B.J. Shea, B.C. Reeves, G. Wells, M. Thuku, C. Hamel, J. Moran, D. Moher, P. Tugwell, V. Welch, E. Kristjansson, D.A. Henry, AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both, BMJ. 358 (2017) 4008. doi:10.1136/BMJ.J4008.
12. E. Mohammadi, M. Thelwall, S. Haustein, V. Larivière, Who reads research articles? An altmetrics analysis of Mendeley user categories, J. Assoc. Inf. Sci. Technol. 66 (2015) 1832–1846. doi:10.1002/asi.23286.
13. A. Sharma, T. Medapalli, M. Alexandrou, E. Brilakis, A. Prasad, Exploring the Role of ChatGPT in Cardiology: A Systematic Review of the Current Literature., Cureus. 16 (2024) e58936. doi:10.7759/cureus.58936.
14. V. Sorin, B.S. Glicksberg, Y. Artsi, Y. Barash, E. Konen, G.N. Nadkarni, E. Klang, Utilizing large language models in breast cancer management: systematic review., J. Cancer Res. Clin. Oncol. 150 (2024) 140. doi:10.1007/s00432-024-05678-6.
15. M. Omar, S. Soffer, A.W. Charney, I. Landi, G.N. Nadkarni, E. Klang, Applications of large language models in psychiatry: a systematic review, Front. Psychiatry. 15 (2024). doi:10.3389/fpsyt.2024.1422807.
16. Z. Guo, A. Lai, J.H. Thygesen, J. Farrington, T. Keen, K. Li, Large Language Models for Mental Health Applications: Systematic Review, JMIR Ment. Heal. 11 (2024). doi:10.2196/57400.
17. M. Omar, I. Levkovich, Exploring the efficacy and potential of large language models for depression: A systematic review, J. Affect. Disord. 371 (2025) 234 – 244. doi:10.1016/j.jad.2024.11.052.
18. H.J. Dai, C.H. Su, Y.Q. Lee, Y.C. Zhang, C.K. Wang, C.J. Kuo, C.S. Wu, Deep Learning-Based Natural Language Processing for Screening Psychiatric Patients, Front. Psychiatry. 11 (2021) 533949. doi:10.3389/FPSYT.2020.533949/BIBTEX.
19. A. Pourkeyvan, R. Safa, A. Sorourkhah, Harnessing the Power of Hugging Face Transformers for Predicting Mental Health Disorders in Social Networks, IEEE Access. 12 (2024) 28025–28035. doi:10.1109/ACCESS.2024.3366653.
20. J.A. Lossio-Ventura, R. Weger, A.Y. Lee, E.P. Guinee, J. Chung, L. Atlas, E. Linos, F. Pereira, A Comparison of ChatGPT and Fine-Tuned Open Pre-Trained Transformers (OPT) Against Widely Used Sentiment Analysis Tools: Sentiment Analysis of COVID-19 Survey Data., JMIR Ment. Heal. 11 (2024) e50150. doi:10.2196/50150.
21. R. Franco D'Souza, S. Amanullah, M. Mathew, K.M. Surapaneni, Appraising the performance of ChatGPT in psychiatry using 100 clinical case vignettes, Asian J. Psychiatr. 89 (2023) 103770. doi:10.1016/J.AJP.2023.103770.
22. C. Lau, X. Zhu, W.Y. Chan, Automatic depression severity assessment with deep learning using parameter-efficient tuning, Front. Psychiatry. 14 (2023) 1160291. doi:10.3389/FPSYT.2023.1160291/BIBTEX.
23. B.G. Bokolo, Q. Liu, Advanced Comparative Analysis of Machine Learning and Transformer Models for Depression and Suicide Detection in Social Media Texts, Electron. 2024, Vol. 13, Page 3980. 13 (2024) 3980. doi:10.3390/ELECTRONICS13203980.
24. Z. Elyoseph, I. Levkovich, S. Shinan-Altman, Assessing prognosis in depression: comparing perspectives of AI models, mental health professionals and the general public, Fam. Med. Community Heal. 12 (2024) e002583. doi:10.1136/FMCH-2023-002583.
25. I. Levkovich, Z. Elyoseph, Identifying depression and its determinants upon initiating treatment: ChatGPT versus primary care physicians, Fam. Med. Community Heal. 11 (2023) e002391. doi:10.1136/FMCH-2023-002391.
26. I. Levkovich, Z. Elyoseph, Suicide Risk Assessments Through the Eyes of ChatGPT-3.5 Versus ChatGPT-4: Vignette Study., JMIR Ment. Heal. 10 (2023) e51232. doi:10.2196/51232.

27. X. Wang, S. Chen, T. Li, W. Li, Y. Zhou, J. Zheng, Q. Chen, J. Yan, B. Tang, Depression Risk Prediction for Chinese Microblogs via Deep-Learning Methods: Content Analysis., JMIR Med. Informatics. 8 (2020) e17958–e17958. doi:10.2196/17958.
28. R.H. Perlis, J.F. Goldberg, M.J. Ostacher, C.D. Schneck, Clinical decision support for bipolar depression using large language models, Neuropsychopharmacology. 49 (2024) 1412–1416. doi:10.1038/S41386-024-01841-2;SUBJMETA.
29. A.N. Vaidyam, H. Wisniewski, J.D. Halamka, M.S. Kashavan, J.B. Torous, Chatbots and Conversational Agents in Mental Health: A Review of the Psychiatric Landscape, Can. J. Psychiatry. 64 (2019) 456–464. doi:10.1177/0706743719828977.
30. S.K. Ahmed, S. Hussein, T.A. Aziz, S. Chakraborty, M.R. Islam, K. Dhama, The power of ChatGPT in revolutionizing rural healthcare delivery, Heal. Sci. Reports. 6 (2023) e1684. doi:10.1002/HSR2.1684.
31. D. Chatziisaak, P. Burri, M. Sparn, D. Hahnloser, T. Steffen, S. Bischofberger, Concordance of ChatGPT artificial intelligence decision-making in colorectal cancer multidisciplinary meetings: retrospective study, BJS Open. 9 (2025). doi:10.1093/BJSOPEN/ZRAF040.
32. D. Gibson, S. Jackson, R. Shanmugasundaram, I. Seth, A. Siu, N. Ahmadi, J. Kam, N. Mehan, R. Thanigasalam, N. Jeffery, M.I. Patel, S. Leslie, Evaluating the Efficacy of ChatGPT as a Patient Education Tool in Prostate Cancer: Multimetric Assessment, J. Med. Internet Res. 26 (2024) e55939. doi:10.2196/55939.
33. A. Bracken, C. Reilly, A. Feeley, E. Sheehan, K. Merghani, I. Feeley, Artificial Intelligence (AI) – Powered Documentation Systems in Healthcare: A Systematic Review, J. Med. Syst. 49 (2025) 1–10. doi:10.1007/S10916-025-02157-4/TABLES/8.
34. S. Banerjee, A. Agarwal, S. Singla, LLMs Will Always Hallucinate, and We Need to Live with This, Lect. Notes Networks Syst. 1554 LNNS (2025) 624–648. doi:10.1007/978-3-031-99965-9_39.
35. J. Jiao, S. Afroogh, Y. Xu, C. Phillips, Navigating LLM ethics: advancements, challenges, and future directions, AI Ethics 2025. (2025) 1–25. doi:10.1007/S43681-025-00814-5.
36. K. Kusunose, S. Kashima, M. Sata, Evaluation of the Accuracy of ChatGPT in Answering Clinical Questions on the Japanese Society of Hypertension Guidelines, Circ. J. 87 (2023) 1030–1033. doi:10.1253/CIRCJ.CJ-23-0308.
37. A. Rizwan, T. Sadiq, The Use of AI in Diagnosing Diseases and Providing Management Plans: A Consultation on Cardiovascular Disorders With ChatGPT, (n.d.). doi:10.7759/cureus.43106.
38. I. Skalidis, A. Cagnina, W. Luangphiphat, T. Mahendiran, O. Muller, E. Abbe, S. Fournier, ChatGPT takes on the European Exam in Core Cardiology: an artificial intelligence success story?, Eur. Hear. Journal. Digit. Heal. 4 (2023) 279. doi:10.1093/EHJDH/ZTAD029.
39. L. Van Bulck, P. Moons, What if your patient switches from Dr. Google to Dr. ChatGPT? A vignette-based survey of the trustworthiness, value, and danger of ChatGPT-generated responses to health questions, Eur. J. Cardiovasc. Nurs. 23 (2024) 95–98. doi:10.1093/EURJCN/ZVAD038.
40. M.C. Williams, J. Shambrook, How will artificial intelligence transform cardiovascular computed tomography? A conversation with an AI model, J. Cardiovasc. Comput. Tomogr. 17 (2023) 281–283. doi:10.1016/J.JCCT.2023.03.010.
41. H.S. Choi, J.Y. Song, K.H. Shin, J.H. Chang, B.S. Jang, Developing prompts from large language model for extracting clinical information from pathology and ultrasound reports in breast cancer, Radiat. Oncol. J. 41 (2023) 209. doi:10.3857/ROJ.2023.00633.
42. S. Griewing, N. Gremke, U. Wagner, M. Lingenfelder, S. Kuhn, J. Boekhoff, Challenging ChatGPT 3.5 in Senology—An Assessment of Concordance with Breast Cancer Tumor Board Decision Making, J. Pers. Med. 13 (2023) 1502. doi:10.3390/JPM13101502/S1.
43. H.L. Haver, E.B. Ambinder, M. Bahl, E.T. Oluyemi, J. Jeudy, P.H. Yi, Appropriateness of Breast Cancer Prevention and Screening Recommendations Provided by ChatGPT, Https://Doi.Org/10.1148/Radiol.230424. 307 (2023). doi:10.1148/RADIOL.230424.
44. S. Lukac, D. Dayan, V. Fink, E. Leinert, A. Hartkopf, K. Veselinovic, W. Janni, B. Rack, K. Pfister, B. Heitmeir, F. Ebner, Evaluating ChatGPT as an adjunct for the multidisciplinary tumor board decision-making in primary breast cancer cases, Arch. Gynecol. Obstet. 308 (2023) 1831–1844. doi:10.1007/S00404-023-07130-5/TABLES/3.

45. A. Rao, J. Kim, M. Kamineni, M. Pang, W. Lie, K.J. Dreyer, M.D. Succi, Evaluating GPT as an Adjunct for Radiologic Decision Making: GPT-4 Versus GPT-3.5 in a Breast Imaging Pilot, J. Am. Coll. Radiol. 20 (2023) 990–997. doi:10.1016/J.JACR.2023.05.003.
46. V. Sorin, E. Klang, M. Sklair-Levy, I. Cohen, D.B. Zippel, N. Balint Lahat, E. Konen, Y. Barash, Large language model (ChatGPT) as a support tool for breast tumor board, Npj Breast Cancer. 9 (2023) 1–4. doi:10.1038/S41523-023-00557-8;SUBJMETA.
47. A. Abilkaiyrkyzy, F. Laamarti, M. Hamdi, A. El Saddik, Dialogue System for Early Mental Illness Detection: Toward a Digital Twin Solution, IEEE Access. 12 (2024) 2007–2024. doi:10.1109/ACCESS.2023.3348783.
48. S. Adarsh, B. Antony, SSN@LT-EDI-ACL2022: Transfer Learning using BERT for Detecting Signs of Depression from Social Media Texts, LTEDI 2022 - 2nd Work. Lang. Technol. Equal. Divers. Inclusion, Proc. Work. (2022) 326–330. doi:10.18653/V1/2022.LTEDI-1.50.
49. A. Alessa, H. Al-Khalifa, Towards Designing a ChatGPT Conversational Companion for Elderly People, ACM Int. Conf. Proceeding Ser. (2023) 667–674. doi:10.1145/3594806.3596572.
50. J.L. Beredo, E.C. Ong, A Hybrid Response Generation Model for an Empathetic Conversational Agent, 2022 Int. Conf. Asian Lang. Process. IALP 2022. (2022) 300–305. doi:10.1109/IALP57159.2022.9961311.
51. S. Berrezueta-Guzman, M. Kandil, M.L. Martín-Ruiz, I. Pau de la Cruz, S. Krusche, Future of ADHD Care: Evaluating the Efficacy of ChatGPT in Therapy Enhancement, Healthc. 2024, Vol. 12, Page 683. 12 (2024) 683. doi:10.3390/HEALTHCARE12060683.
52. C. Blease, A. Worthen, J. Torous, Psychiatrists' experiences and opinions of generative artificial intelligence in mental healthcare: An online mixed methods survey, Psychiatry Res. 333 (2024) 115724. doi:10.1016/J.PSYCHRES.2024.115724.
53. R. Crasto, L. Dias, D. Miranda, D. Kayande, CareBot: A mental health chatbot, 2021 2nd Int. Conf. Emerg. Technol. INCET 2021. (2021). doi:10.1109/INCET51464.2021.9456326.
54. M. Danner, B. Hadzic, S. Gerhardt, S. Ludwig, I. Uslu, P. Shao, T. Weber, Y. Shiban, M. Rätsch, Advancing Mental Health Diagnostics: GPT-Based Method for Depression Detection, 2023 62nd Annu. Conf. Soc. Instrum. Control Eng. SICE 2023. (2023) 1290–1296. doi:10.23919/SICE59929.2023.10354236.
55. I. Dergaa, F. Fekih-Romdhane, S. Hallit, A.A. Loch, J.M. Glenn, M.S. Fessi, M. Ben Aissa, N. Souissi, N. Guelmami, S. Swed, A. El Omri, N.L. Bragazzi, H. Ben Saad, ChatGPT is not ready yet for use in providing mental health assessment and interventions, Front. Psychiatry. 14 (2023) 1277756. doi:10.3389/FPSYT.2023.1277756/BIBTEX.
56. E.J.S. Diniz, J.E. Fontenele, A.C. de Oliveira, V.H. Bastos, S. Teixeira, R.L. Rabêlo, D.B. Calçada, R.M. Dos Santos, A.K. de Oliveira, A.S. Teles, Boamente: A Natural Language Processing-Based Digital Phenotyping Tool for Smart Monitoring of Suicidal Ideation, Healthc. 2022, Vol. 10, Page 698. 10 (2022) 698. doi:10.3390/HEALTHCARE10040698.
57. Z. Elyoseph, D. Hadar-Shoval, K. Asraf, M. Lvovsky, ChatGPT outperforms humans in emotional awareness evaluations, Front. Psychol. 14 (2023) 1199058. doi:10.3389/FPSYG.2023.1199058/BIBTEX.
58. Z. Elyoseph, I. Levkovich, Beyond human expertise: the promise and limitations of ChatGPT in suicide risk assessment, Front. Psychiatry. 14 (2023) 1213141. doi:10.3389/FPSYT.2023.1213141/BIBTEX.
59. S. Esackimuthu, H. Shruthi, R. Sivanaiah, S. Angel Deborah, R. Sakaya Milton, T.T. Mirnalinee, SSN_MLRG3 @LT-EDI-ACL2022-Depression Detection System from Social Media Text using Transformer Models, LTEDI 2022 - 2nd Work. Lang. Technol. Equal. Divers. Inclusion, Proc. Work. (2022) 196–199. doi:10.18653/V1/2022.LTEDI-1.26.
60. F. Farhat, ChatGPT as a Complementary Mental Health Resource: A Boon or a Bane, Ann. Biomed. Eng. 52 (2024) 1111–1114. doi:10.1007/S10439-023-03326-7/METRICS.
61. N. Farruque, O.R. Zaïane, R. Goebel, S. Sivapalan, DeepBlues@LT-EDI-ACL2022: Depression level detection modelling through domain specific BERT and short text Depression classifiers, LTEDI 2022 - 2nd Work. Lang. Technol. Equal. Divers. Inclusion, Proc. Work. (2022) 167–171. doi:10.18653/V1/2022.LTEDI-1.21.
62. N. Farruque, R. Goebel, S. Sivapalan, O.R. Zaïane, Depression symptoms modelling from social media text: an LLM driven semi-supervised learning approach, Lang. Resour. Eval. 58 (2024) 1013–1041. doi:10.1007/S10579-024-09720-4/TABLES/23.

 doi:10.20944/preprints202510.2480.v1

63. S.F. Friedman, G. Ballentine, Trajectories of sentiment in 11,816 psychoactive narratives, Hum. Psychopharmacol. Clin. Exp. 39 (2024) e2889. doi:10.1002/HUP.2889.
64. H. Ghanadian, I. Nejadgholi, H. Al Osman, Socially Aware Synthetic Data Generation for Suicidal Ideation Detection Using Large Language Models, IEEE Access. 12 (2024) 14350–14363. doi:10.1109/ACCESS.2024.3358206.
65. D. Hadar-Shoval, Z. Elyoseph, M. Lvovsky, The plasticity of ChatGPT's mentalizing abilities: personalization for personality structures, Front. Psychiatry. 14 (2023) 1234397. doi:10.3389/FPSYT.2023.1234397/BIBTEX.
66. M.F.M. Hayati, M.A.M. Ali, A.N.M. Rosli, Depression Detection on Malay Dialects Using GPT-3, 7th IEEE-EMBS Conf. Biomed. Eng. Sci. IECBES 2022 - Proc. (2022) 360–364. doi:10.1109/IECBES54088.2022.10079554.
67. W. He, W. Zhang, Y. Jin, Q. Zhou, H. Zhang, Q. Xia, Physician Versus Large Language Model Chatbot Responses to Web-Based Questions From Autistic Patients in Chinese: Cross-Sectional Comparative Analysis, J. Med. Internet Res. 26 (2024) e54706. doi:10.2196/54706.
68. A. Hegde, S. Coelho, A.E. Dashti, H.L. Shashirekha, MUCS@Text-LT-EDI@ACL 2022: Detecting Sign of Depression from Social Media Text using Supervised Learning Approach, LTEDI 2022 - 2nd Work. Lang. Technol. Equal. Divers. Inclusion, Proc. Work. (2022) 312–316. doi:10.18653/V1/2022.LTEDI-1.47.
69. T.F. Heston, Safety of Large Language Models in Addressing Depression, Cureus. 15 (2023) e50729. doi:10.7759/CUREUS.50729.
70. A. de Hond, M. van Buchem, C. Fanconi, M. Roy, D. Blayney, I. Kant, E. Steyerberg, T. Hernandez-Boussard, Predicting Depression Risk in Patients With Cancer Using Multimodal Data: Algorithm Development Study., JMIR Med. Informatics. 12 (2024) e51925. doi:10.2196/51925.
71. D. Howard, M.M. Maslej, J. Lee, J. Ritchie, G. Woollard, L. French, Transfer learning for risk classification of social media posts: Model evaluation study, J. Med. Internet Res. 22 (2020) e15371. doi:10.2196/15371.
72. G. Hwang, D.Y. Lee, S. Seol, J. Jung, Y. Choi, E.S. Her, M.H. An, R.W. Park, Assessing the potential of ChatGPT for psychodynamic formulations in psychiatry: An exploratory study, Psychiatry Res. 331 (2024) 115655. doi:10.1016/J.PSYCHRES.2023.115655.
73. L. Ilias, S. Mouzakitis, D. Askounis, Calibration of Transformer-Based Models for Identifying Stress and Depression in Social Media, IEEE Trans. Comput. Soc. Syst. 11 (2024) 1979–1990. doi:10.1109/TCSS.2023.3283009.
74. M. Janatdoust, F. Ehsani-Besheli, H. Zeinali, KADO@LT-EDI-ACL2022: BERT-based Ensembles for Detecting Signs of Depression from Social Media Text, LTEDI 2022 - 2nd Work. Lang. Technol. Equal. Divers. Inclusion, Proc. Work. (2022) 265–269. doi:10.18653/V1/2022.LTEDI-1.38.
75. M. Kabir, T. Ahmed, M.B. Hasan, M.T.R. Laskar, T.K. Joarder, H. Mahmud, K. Hasan, DEPTWEET: A typology for social media texts to detect depression severities, Comput. Human Behav. 139 (2023) 107503. doi:10.1016/J.CHB.2022.107503.
76. H. Kumar, Y. Wang, J. Shi, I. Musabirov, N.A.S. Farb, J.J. Williams, Exploring the Use of Large Language Models for Improving the Awareness of Mindfulness, Conf. Hum. Factors Comput. Syst. - Proc. 7 (2023). doi:10.1145/3544549.3585614/SUPPL_FILE/3544549.3585614-TALK-VIDEO.MP4.
77. G. Lam, H. Dongyan, W. Lin, Context-aware Deep Learning for Multi-modal Depression Detection, ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc. 2019-May (2019) 3946–3950. doi:10.1109/ICASSP.2019.8683027.
78. D.J. Li, Y.C. Kao, S.J. Tsai, Y.M. Bai, T.C. Yeh, C.S. Chu, C.W. Hsu, S.W. Cheng, T.W. Hsu, C.S. Liang, K.P. Su, Comparing the performance of ChatGPT GPT-4, Bard, and Llama-2 in the Taiwan Psychiatric Licensing Examination and in differential diagnosis with multi-center psychiatrists, Psychiatry Clin. Neurosci. 78 (2024) 347–352. doi:10.1111/PCN.13656.
79. C. Liyanage, M. Garg, V. Mago, S. Sohn, Augmenting Reddit Posts to Determine Wellness Dimensions impacting Mental Health, Proc. Conf. Assoc. Comput. Linguist. Meet. 2023 (2023) 306. doi:10.18653/V1/2023.BIONLP-1.27.
80. K.C. Lu, S.A. Thamrin, A.L.P. Chen, Depression detection via conversation turn classification, Multimed. Tools Appl. 82 (2023) 39393–39413. doi:10.1007/S11042-023-15103-8/TABLES/13.

81. Z. Ma, Y. Mei, Z. Su, Understanding the Benefits and Challenges of Using Large Language Model-based Conversational Agents for Mental Well-being Support, AMIA Annu. Symp. Proc. 2023 (2024) 1105. https://pmc.ncbi.nlm.nih.gov/articles/PMC10785945/ (accessed October 17, 2025).
82. H. Mazumdar, C. Chakraborty, M. Sathvik, sabyasachi Mukhopadhyay, P.K. Panigrahi, GPTFX: A Novel GPT-3 Based Framework for Mental Health Detection and Explanations, IEEE J. Biomed. Heal. Informatics. (2023). doi:10.1109/JBHI.2023.3328350.
83. H. Metzler, H. Baginski, T. Niederkrotenthaler, D. Garcia, Detecting Potentially Harmful and Protective Suicide-Related Content on Twitter: Machine Learning Approach, J. Med. Internet Res. 24 (2022) e34705. doi:10.2196/34705.
84. D. Owen, D. Antypas, A. Hassoulas, A.F. Pardiñas, L. Espinosa-Anke, J.C. Collados, Enabling Early Health Care Intervention by Detecting Depression in Users of Web-Based Forums using Language Models: Longitudinal Analysis and Evaluation., JMIR AI. 2 (2023) e41205. doi:10.2196/41205.
85. G. Parker, M.J. Spoelma, A chat about bipolar disorder, Bipolar Disord. 26 (2024) 249–254. doi:10.1111/BDI.13379.
86. R. Poswiata, M. Perełkiewicz, OPI@LT-EDI-ACL2022: Detecting Signs of Depression from Social Media Text using RoBERTa Pre-trained Language Models, LTEDI 2022 - 2nd Work. Lang. Technol. Equal. Divers. Inclusion, Proc. Work. (2022) 276–282. doi:10.18653/V1/2022.LTEDI-1.40.
87. M. Sadeghi, B. Egger, R. Agahi, R. Richer, K. Capito, L.H. Rupp, L. Schindler-Gmelch, M. Berking, B.M. Eskofier, Exploring the Capabilities of a Language Model-Only Approach for Depression Detection in Text Data, BHI 2023 - IEEE-EMBS Int. Conf. Biomed. Heal. Informatics, Proc. (2023). doi:10.1109/BHI58575.2023.10313367.
88. M.C. Schubert, W. Wick, V. Venkataramani, Performance of Large Language Models on a Neurology Board–Style Examination, JAMA Netw. Open. 6 (2023) e2346721–e2346721. doi:10.1001/JAMANETWORKOPEN.2023.46721.
89. S. Senn, M.L. Tlachac, R. Flores, E. Rundensteiner, Ensembles of BERT for Depression Classification., Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. IEEE Eng. Med. Biol. Soc. Annu. Int. Conf. 2022 (2022) 4691–4694. doi:10.1109/EMBC48229.2022.9871120.
90. M. Singh, P. Motlicek, IDIAP Submission@LT-EDI-ACL2022: Detecting Signs of Depression from Social Media Text, Proc. Second Work. Lang. Technol. Equal. Divers. Incl. (2022) 362–368. doi:10.18653/V1/2022.LTEDI-1.56.
91. S. Sivamanikandan, V. Santhosh, N. Sanjaykumar, C. Jerin Mahibha, D. Thenmozhi, scubeMSEC@LT-EDI-ACL2022: Detection of Depression using Transformer Models, LTEDI 2022 - 2nd Work. Lang. Technol. Equal. Divers. Inclusion, Proc. Work. (2022) 212–217. doi:10.18653/V1/2022.LTEDI-1.29.
92. S. Spallek, L. Birrell, S. Kershaw, E.K. Devine, L. Thornton, Can we use ChatGPT for Mental Health and Substance Use Education? Examining Its Quality and Potential Harms, JMIR Med. Educ. 9 (2023) e51243. doi:10.2196/51243.
93. W. Stigall, M.A. Al Hafiz Khan, D. Attota, F. Nweke, Y. Pei, Large Language Models Performance Comparison of Emotion and Sentiment Classification, Proc. 2024 ACM Southeast Conf. ACMSE 2024. (2024) 60–68. doi:10.1145/3603287.3651183.
94. M. Suri, N. Semwal, D. Chaudhary, I. Gorton, B. Kumar, I don't feel so good! Detecting Depressive Tendencies using Transformer-based Multimodal Frameworks, ACM Int. Conf. Proceeding Ser. (2022) 360–365. doi:10.1145/3578741.3578817.
95. Y. Tao, M. Yang, H. Shen, Z. Yang, Z. Weng, B. Hu, Classifying Anxiety and Depression through LLMs Virtual Interactions: A Case Study with ChatGPT, Proc. - 2023 2023 IEEE Int. Conf. Bioinforma. Biomed. BIBM 2023. (2023) 2259–2264. doi:10.1109/BIBM58861.2023.10385305.
96. W.L. Tey, H.N. Goh, A.H.L. Lim, C.K. Phang, Pre- and Post-Depressive Detection using Deep Learning and Textual-based Features, Int. J. Technol. 14 (2023) 1334–1343. doi:10.14716/IJTECH.V14I6.6648.
97. E. Toto, M.L. Tlachac, E.A. Rundensteiner, AudiBERT: A Deep Transfer Learning Multimodal Classification Framework for Depression Screening, Int. Conf. Inf. Knowl. Manag. Proc. (2021) 4145–4154. doi:10.1145/3459637.3481895/SUPPL_FILE/CIKM21-AFP0975.MP4.

 doi:10.20944/preprints202510.2480.v1

98. V. Vajre, M. Naylor, U. Kamath, A. Shehu, PsychBERT: A Mental Health Language Model for Social Media Mental Health Behavioral Analysis, Proc. - 2021 IEEE Int. Conf. Bioinforma. Biomed. BIBM 2021. (2021) 1077–1082. doi:10.1109/BIBM52615.2021.9669469.
99. S. Verma, Vishal, R.C. Joshi, M.K. Dutta, S. Jezek, R. Burget, AI-Enhanced Mental Health Diagnosis: Leveraging Transformers for Early Detection of Depression Tendency in Textual Data, Int. Congr. Ultra Mod. Telecommun. Control Syst. Work. (2023) 56–61. doi:10.1109/ICUMT61075.2023.10333301.
100. C. Wan, X. Ge, J. Wang, X. Zhang, Y. Yu, J. Hu, Y. Liu, H. Ma, Identification and Impact Analysis of Family History of Psychiatric Disorder in Mood Disorder Patients With Pretrained Language Model, Front. Psychiatry. 13 (2022) 861930. doi:10.3389/FPSYT.2022.861930/BIBTEX.
101. X. Wang, K. Liu, C. Wang, Knowledge-enhanced Pre-Training large language model for depression diagnosis and treatment, Proceeding 2023 9th IEEE Int. Conf. Cloud Comput. Intell. Syst. CCIS 2023. (2023) 532–536. doi:10.1109/CCIS59572.2023.10263217.
102. Y. Wei, L. Guo, C. Lian, J. Chen, ChatGPT: Opportunities, risks and priorities for psychiatry, Asian J. Psychiatr. 90 (2023) 103808. doi:10.1016/J.AJP.2023.103808.
103. Y. Wu, J. Chen, K. Mao, Y. Zhang, Automatic Post-Traumatic Stress Disorder Diagnosis via Clinical Transcripts: A Novel Text Augmentation with Large Language Models, BioCAS 2023 - 2023 IEEE Biomed. Circuits Syst. Conf. Conf. Proc. (2023). doi:10.1109/BIOCAS58349.2023.10388714.
104. N. Yongsatianchot, P.G. Torshizi, S. Marsella, Investigating Large Language Models' Perception of Emotion Using Appraisal Theory, 2023 11th Int. Conf. Affect. Comput. Intell. Interact. Work. Demos. (2023) 1–8. doi:10.1109/ACIIW59127.2023.10388194.
105. Y. Zhang, H. Lyu, Y. Liu, X. Zhang, Y. Wang, J. Luo, Monitoring Depression Trends on Twitter During the COVID-19 Pandemic: Observational Study., JMIR Infodemiology. 1 (2021) e26769–e26769. doi:10.2196/26769.