

International Journal of Innovative Technologies in Social Science

e-ISSN: 2544-9435

Operating Publisher
SciFormat Publishing Inc.
ISNI: 0000 0005 1449 8214

2734 17 Avenue SW,
Calgary, Alberta, T3E0A7,
Canada
+15878858911
editorial-office@sciformat.ca

ARTICLE TITLE	LARGE LANGUAGE MODELS IN MENTAL HEALTH TRIAGE: A SYSTEMATIC REVIEW OF ACCURACY, BIAS, AND ETHICAL DEPLOYMENT ACROSS DIGITAL PLATFORMS
DOI	https://doi.org/10.31435/ijitss.4(52).2026.6752
RECEIVED	22 August 2026
ACCEPTED	17 October 2026
PUBLISHED	26 October 2026
LICENSE	The article is licensed under a **Creative Commons Attribution 4.0 International License.**

© The author(s) 2026.
This article is published as open access under the Creative Commons Attribution 4.0 International License (CC BY 4.0), allowing the author to retain copyright. The CC BY 4.0 License permits the content to be copied, adapted, displayed, distributed, republished, or reused for any purpose, including adaptation and commercial use, as long as proper attribution is provided.

LARGE LANGUAGE MODELS IN MENTAL HEALTH TRIAGE: A SYSTEMATIC REVIEW OF ACCURACY, BIAS, AND ETHICAL DEPLOYMENT ACROSS DIGITAL PLATFORMS

Alicja Kwiatkowska *(Corresponding Author, Email: alicja.kwiatkowska75@gmail.com)*
Pomeranian Medical University (Pomorski Uniwersytet Medyczyy), Szczecin, Poland
ORCID ID: 0009-0003-1027-2114

Mateusz Lis
Pomeranian Medical University (Pomorski Uniwersytet Medyczny), Szczecin, Poland
ORCID ID: 0009-0006-8441-8296

Zuzanna Julia Komin
Pomeranian Medical University (Pomorski Uniwersytet Medyczny), Szczecin, Poland
ORCID ID: 0009-0001-6284-6410

Maryan Liasota
Pomeranian Medical University (Pomorski Uniwersytet Medyczny), Szczecin, Poland
ORCID ID: 0009-0006-0266-4849

Eliza Zofia Kabat
Pomeranian Medical University (Pomorski Uniwersytet Medyczny), Szczecin, Poland
ORCID ID: 0009-0003-6531-967X

Michalina Justyna Kozłowska
Pomeranian Medical University (Pomorski Uniwersytet Medyczny), Szczecin, Poland
ORCID ID: 0009-0003-9561-2957

Krzysztof Konopka
Pomeranian Medical University (Pomorski Uniwersytet Medyczny), Szczecin, Poland
ORCID ID: 0009-0003-4055-0074

Maciej Kamil Majchrzak
Pomeranian Medical University (Pomorski Uniwersytet Medyczny), Szczecin, Poland
ORCID ID: 0009-0007-4739-0092

Faustyna Magdalena Kołodziej
Pomeranian Medical University (Pomorski Uniwersytet Medyczny), Szczecin, Poland
ORCID ID: 0009-0002-8778-7100

Jakub Juskowiak
Poznań University of Medical Sciences, Poznań, Poland
ORCID ID: 0009-0001-0372-400X

ABSTRACT

Background: The intersection of a global psychiatric workforce shortage and the rapid advancement of Generative Artificial Intelligence (GenAI) has catalyzed the integration of Large Language Models (LLMs) into digital healthcare. LLMs are increasingly deployed as preliminary triage agents to assess psychological distress, yet their diagnostic reliability, demographic neutrality, and safety protocols in crisis scenarios remain heavily debated within the psychiatric community.
Objective: This review evaluates current research to determine the trustworthiness of large language models (LLMs) used in psychiatric screening. In doing so, it shines a light on the algorithmic biases present in practical use and maps out the necessary ethical standards to ensure digital health services remain safe and fair for all patients.
Methodology: We sourced peer-reviewed publications released between 2023 and 2026 by running a systematic literature search across Google Scholar, PubMed/MEDLINE, and IEEE Xplore. The synthesis focused on comparative accuracy studies, observational evaluations, and sociotechnical analyses of leading LLMs (e.g., GPT-4, Claude 3.5 Sonnet, Gemini 1.5 Pro) applied specifically to psychiatric screening and automated clinical routing.
Results: LLMs demonstrate high diagnostic concordance with human clinicians in structured assessments for high-prevalence conditions like major depressive disorder, effectively parsing unstructured clinical narratives. However, they exhibit a consistent tendency toward conservative over-triage in psychiatric emergencies. Profound algorithmic biases persist; models frequently recommend divergent, often inferior, treatment pathways for marginalized demographic groups and display gender-based stigmatization. Furthermore, models struggle with complex psychopathologies and display inconsistent safety mechanisms during acute suicidal ideation queries.
Conclusion: While LLMs offer substantial utility for scaling preliminary psychiatric screening and reducing administrative triage burdens, their propensity for factual hallucination, demographic bias, and uncalibrated risk escalation currently precludes autonomous clinical deployment. Robust, human-in-the-loop governance and culturally representative training corpora are imperative prior to further integration into global digital health ecosystems.

KEYWORDS

Large Language Models, Mental Health Triage, Algorithmic Bias, Psychiatric Screening, Digital Health Equity, Generative Artificial Intelligence

CITATION

Kwiatkowska, A., Lis, M., Komin, Z. J., Liasota, M., Kabat, E. Z., Kozłowska, M. J., Konopka, K., Majchrzak, M. K., Kołodziej, F. M., & Juskowiak, J. (2026). Large language models in mental health triage: A systematic review of accuracy, bias, and ethical deployment across digital platforms. *International Journal of Innovative Technologies in Social Science*, 4(52). https://doi.org/10.31435/ijitss.4(52).2026.6752

COPYRIGHT

© The author(s) 2026. This article is published as open access under the **Creative Commons Attribution 4.0 International License (CC BY 4.0)**, allowing the author to retain copyright. The CC BY 4.0 License permits the content to be copied, adapted, displayed, distributed, republished, or reused for any purpose, including adaptation and commercial use, as long as proper attribution is provided.

1. Introduction

1.1 Background: The Mental Health Triage Crisis

Medical systems worldwide are currently buckling under a severe psychiatric emergency. Hospitals and clinics face an impossible balancing act: patient numbers are soaring, cases are growing more complicated, and there simply are not enough trained professionals to help. Research shows that 50% of adults will face some form of mental illness as they age. Tragically, almost 90% of those battling the most severe disorders go without proper treatment. Wealthy countries are not immune to this breakdown. Hamstrung by inefficient frameworks and a worldwide deficit of roughly 5 million clinicians, wait lists have exploded. For some groups of patients, an average of eleven years passes from the day they first notice symptoms until they finally receive care.

In advanced health systems, such as the United Kingdom's National Health Service (NHS), between 370,000 and 470,000 new referrals for secondary mental healthcare are generated monthly, creating an insurmountable administrative and clinical triage bottleneck (Taylor et al., 2024). Triage, the critical initial assessment determining the acuity of a patient's psychiatric state and the appropriate locus of care, requires synthesizing complex, unstructured narrative data from Electronic Health Records (EHRs) and patient self-reports. The chronic strain on clinical resources often compromises the consistency and speed of these triage decisions, leaving highly vulnerable populations – particularly adolescents and individuals experiencing acute

psychosocial crises – without timely psychiatric intervention. Consequently, health systems are aggressively pursuing scalable digital solutions to augment clinical workflows, expedite preliminary screening, and democratize access to basic psychological support.

1.2 The Emergence of LLMs in Psychiatric Assessment

To bridge the substantial gap between psychiatric demand and resource availability, artificial intelligence (AI) has rapidly transitioned from deterministic, rule-based chatbots (e.g., ELIZA or early algorithmic cognitive behavioral therapy applications) to highly sophisticated Generative AI and Large Language Models (LLMs). Built upon the Transformer architecture and optimized through massive pre-training, supervised fine-tuning (SFT), and reinforcement learning from human feedback (RLHF), models such as OpenAI's GPT-4, Google's Gemini, Anthropic's Claude, and specialized clinical variants possess advanced natural language understanding capabilities that mimic human clinical reasoning.

These models are currently being investigated and deployed across the entire psychiatric care continuum, encompassing automated outpatient reception, digital phenotyping, symptom summarization, and acute risk stratification. Unlike traditional psychometric scales, LLMs excel at processing unstructured free-text clinical notes, interpreting nuanced emotional language, and engaging patients in dynamic, empathetic conversational assessments. Recent trials demonstrating their capability to pass medical board examinations with high proficiency in psychiatry have spurred aggressive integration into digital health platforms, positioning LLMs not merely as administrative aids, but as active socio-technical diagnostic agents. For example, customized LLM frameworks analyzing EHRs have demonstrated the capacity to reduce manual psychiatric triage time per case from 10 minutes to 8 minutes, translating to the conservation of 329 person-weeks of human resources per month in specific institutional deployments. (Taylor et al., 2024)

1.3 From Diagnostic Accuracy to Bias and Ethics

Despite their remarkable computational proficiency and syntactic fluency, the probabilistic nature of LLMs introduces novel and severe risks to clinical safety. Because these models generate responses by predicting subsequent token probabilities based on vast, uncurated internet training corpora, they inherently encode, reflect, and replicate the historical prejudices, structural inequities, and clinical inaccuracies embedded within those datasets. In the context of mental health triage, where patient safety relies on the precise detection of subtle linguistic markers indicative of suicidality, severe distress, or psychosis, the stakes of algorithmic error are profoundly high.

The primary operational concern is the phenomenon of "hallucination" – the confident generation of factually incorrect, biologically implausible, or clinically contraindicated information. In acute psychiatric triage, an intrinsic or factual hallucination could manifest as a failure to recognize a patient's suicidal intent, the provision of incorrect pharmacological information, or the generation of narratives that validate and reinforce a patient's pre-existing psychotic framework. Furthermore, an emerging and robust body of evidence highlights profound algorithmic bias. Research indicates that LLMs perform disparately across demographic lines, exhibiting reduced diagnostic accuracy and recommending inferior therapeutic interventions when presented with vignettes involving marginalized racial groups, specific genders, or non-standard linguistic dialects such as African American Vernacular English (AAVE). This raises acute ethical concerns regarding "epistemic violence" – where the subjective distress of marginalized populations is miscategorized, invalidated, or pathologized by systems calibrated predominantly on Western, Educated, Industrialized, Rich, and Democratic (WEIRD) datasets.

1.4 Research Gap and Rationale

The scientific literature surrounding LLMs in mental health is expanding exponentially; however, it remains highly fragmented, methodologically inconsistent, and often isolated within specific computer science or clinical silos. Previous systematic reviews have predominantly evaluated the technical feasibility of models on narrow, non-clinical benchmarks or focused exclusively on non-generative, traditional machine learning models. A critical gap exists regarding the comprehensive, interdisciplinary synthesis of how contemporary LLMs perform in realistic clinical triage scenarios, specifically concerning the intersection of their diagnostic accuracy with their demographic fairness and ethical safety mechanisms. Furthermore, many empirical evaluations rely heavily on single-session text prompts utilizing simulated patient vignettes, lacking the longitudinal validation necessary to establish robust medico-legal governance and standardized clinical guidelines.

1.5 Aim of the Review

This paper takes a critical look at the evidence surrounding large language models in psychiatric triage and mental health screening. Three core questions guide the analysis. Can LLMs match human practitioners in diagnostic accuracy and risk detection without compromising safety? In what ways do systemic biases within these tools skew care and worsen demographic disparities? And lastly, what regulatory and governance safeguards are truly essential to stop hallucinations, privacy breaches, and botched crisis responses before automated psychiatry is rolled out?

2. Methodology

2.1 Research Design

To ensure methodological rigor, transparency, and reproducibility, this qualitative synthesis was conducted in alignment with the core principles of the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) framework, specifically adapted for scoping and systematic reviews (PRISMA-ScR). The evaluation was structured to capture a holistic, socio-technical understanding of LLMs, viewing the technology not merely as a neutral computational tool, but as an active socio-technical variable directly influencing clinical outcomes, health equity, and medical ethics.

2.2 Data Source and Search Strategy

A comprehensive and systematic literature search was executed utilizing primary academic and medical databases, predominantly PubMed/MEDLINE, with supplementary, peer-reviewed literature sourced from Google Scholar, IEEE Xplore, and the ACM Digital Library. The search was explicitly restricted to the publication window of 2023 to 2026 to ensure the capture of contemporary research evaluating advanced, transformer-based generative AI models (post-ChatGPT emergence) rather than obsolete rule-based conversational agents.

The search strategy utilized a combination of Medical Subject Headings (MeSH) and free-text keywords, employing Boolean operators across targeted domains: ("Large Language Model" OR "LLM" OR "ChatGPT" OR "Generative AI") AND ("Mental Health" OR "Psychiatry" OR "Psychiatric" OR "Psychological") AND ("Triage" OR "Screening" OR "Assessment" OR "Diagnosis") AND ("Accuracy" OR "Bias" OR "Ethics" OR "Safety" OR "Equity")

2.3 Inclusion and Exclusion Criteria

To maintain the highest standards of evidence, studies were screened against predefined criteria. The specific inclusion and exclusion parameters are detailed in Table 1.

Table 1. Study Inclusion and Exclusion Criteria

Criterion	Inclusion Criteria	Exclusion Criteria
Study Type	Peer-reviewed empirical studies, randomized controlled trials (RCTs), systematic reviews, scoping reviews, comparative accuracy studies.	Editorials, opinion pieces, non-peer-reviewed pre-prints without verified data, conference abstracts, and non-academic literature.
Focus	Evaluation of transformer-based LLMs or Generative AI used explicitly for mental health screening, psychiatric triage, or clinical decision support.	Studies evaluating pre-LLM rule-based chatbots, simple decision trees, or generalized AI without a distinct psychiatric triage component.
Database	Indexed in PubMed/MEDLINE, Google Scholar, Scopus, or IEEE Xplore.	Non-indexed sources or proprietary corporate white papers lacking transparent methodological peer review.
Language	Full text available in English.	Articles published in languages other than English (to avoid cross-linguistic translation artifacts in NLP evaluation).
Timeline	Published between January 1, 2023, and June 2026.	Studies published prior to 2023 (pre-dating the widespread clinical evaluation of advanced LLMs).

2.4 Study Selection Process

Our search started with a first pass over titles and abstracts to collect all potentially relevant papers. Duplicate records were removed using reference management software, and the remaining corpus was independently evaluated against the strict inclusion and exclusion parameters by the reviewing team. Studies that evaluated natural language processing (NLP) purely for administrative billing, general medicine, or physical health diagnostics without a clear psychiatric, psychological, or triage-related outcome metric were excluded. A final corpus of highly relevant, high-quality peer-reviewed articles was retained for deep thematic synthesis, ensuring a balanced representation across the domains of clinical accuracy, bias evaluation, and bioethics.

2.5 Data Extraction

Data were systematically extracted into a standardized matrix to capture multiple dimensions of LLM performance. For diagnostic accuracy and safety, extracted variables included the specific LLM architecture evaluated (e.g., GPT-3.5, GPT-4, Claude 3.5 Sonnet, Gemini 1.5 Pro, LLaMA, Meditron), the comparator baseline (e.g., novice clinicians, expert psychiatrists, validated psychometric scales), and the statistical outcome measures (e.g., F1 scores, Cohen's Kappa, Area Under the Curve [AUC], false-positive/negative rates). For bias and equity, extraction focused on the demographic variables tested (race, gender, linguistic dialect, socioeconomic status) and the resulting discrepancies in diagnostic labeling or therapeutic recommendations. Under the ethical domain, attention was directed toward recording hallucination frequencies, patient privacy standards, and crisis escalation procedures.

2.6 Thematic Synthesis

Due to the highly heterogeneous nature of the study designs – ranging from simulated clinical vignette evaluations and prompt-based experiments to retrospective analyses of electronic health records and social media text corpora – a quantitative meta-analysis was deemed methodologically inappropriate. Instead, a narrative thematic synthesis was employed. The extracted evidence was synthesized into three core domains corresponding to the research questions: clinical accuracy, safety and clinician concordance; algorithmic bias and demographic inequities in LLM Triage; and ethical deployment: privacy, liability and crisis management.

This socio-technical framework enables a nuanced evaluation of how technological capabilities intersect with clinical reality and health equity, moving beyond pure computational metrics to assess real-world viability.

3. Results

3.1 Structural Overview of the Evidence

The synthesized evidence demonstrates an explosive, global growth in the application of LLMs across the psychiatric continuum, predominantly driven by the capabilities of the GPT series (GPT-3.5 and GPT-4), alongside rigorous evaluations of Claude, Gemini, and bespoke, locally deployed clinical variants. Recent exhaustive reviews, such as a 2026 synthesis of 205 LLM studies, indicate that the literature overwhelmingly focuses on high-prevalence conditions, notably major depressive disorder, generalized anxiety disorder, and suicidality.

By contrast, far less work has examined severe mental illnesses (SMI), including chronic schizophrenia, bipolar disorder, and co-occurring neurocognitive conditions. This disparity reflects the models' current architectural limitations in processing highly heterogeneous, disorganized, or atypical clinical presentations. The majority of evaluations utilize cross-sectional designs, zero-shot or few-shot prompt experiments using simulated clinical vignettes, or retrospective classification of text corpora (e.g., Reddit posts, EHR notes), with a marked and concerning paucity of longitudinal, real-world randomized controlled trials (RCTs) assessing actual patient outcomes over time.

3.2 Evaluating Diagnostic Performance and Clinician Alignment

Deploying LLMs at the psychiatric intake desk hinges on a simple requirement: the models must reliably replicate a trained physician's clinical judgment. The evidence reveals a bifurcated performance profile: LLMs exhibit remarkable, sometimes superhuman, proficiency in structured, high-prevalence screening tasks, yet falter significantly in complex, ambiguous psychiatric emergencies where contextual nuance is paramount.

In routine screening environments, automated assessment paradigms leveraging LLMs show substantial promise. For example, Jin et al. (2026) developed a paradigm integrating the Beck Depression Inventory Fast Screen with a customized ChatGPT interface (BDI-FS-GPT). During trials, Jin's interface caught roughly nine in ten depression cases (89.3%) and kept false positives down to 11.5%, yielding an AUC of 0.953. This demonstrated a substantial diagnostic agreement (Cohen's $\kappa = 0.72$) with licensed psychiatrists, outperforming traditional static questionnaires.

Similarly, the integration of Retrieval-Augmented Generation (RAG) – a technique that tethers the probabilistic LLM to a deterministic, evidence-based clinical knowledge base – significantly improves diagnostic fidelity and reduces hallucination. An evaluation of RAG-enhanced Chinese LLMs (such as Qwen3 and DeepSeek) achieved an impressive F1-score of 0.9107 for depression detection and significantly enhanced suicide risk stratification compared to non-augmented baselines.

However, in emergency psychiatric triage, LLMs demonstrate a consistent, and potentially problematic, tendency toward uncalibrated conservatism and systemic over-triage. Thotapalli et al. (2025) evaluated three GPT-4 variants (GPT-4o, GPT-4o Mini, GPT-4 Legacy) using 22 acute psychiatric emergency vignettes representing real-world youth mental health crises. While the models achieved substantial interrater reliability with clinicians (Cohen's $\kappa = 0.77$ for admission decisions, 0.78 for risk assessment), they exhibited a clear bias toward overestimating risk, generating zero false negatives but multiple false positives for hospital admission.

We see this same tendency to over-triage in tools designed for social media screening. Settanni et al. (2025) applied GPT-4o, Claude 3.5 Sonnet, and Gemini 1.5 Pro to triage expressions of psychological distress within a corpus of Italian-language Reddit posts. While GPT-4o achieved the highest overall classification performance, all models consistently overestimated the urgency of intervention compared to trained clinicians utilizing the validated Mental Health Triage Scale (MHTS). Claude 3.5 Sonnet, in particular, demonstrated high sensitivity but lower precision, yielding excessive false positives and signaling a tendency to overreact to benign expressions of distress.

Performance also drops off significantly whenever these tools run into rare or clinically complicated disorders. Levkovich (2025) conducted a comparative analysis and found that while ChatGPT-4 matched or slightly exceeded professional accuracy for depression and PTSD vignettes, its performance dropped precipitously to a mere 55% accuracy when attempting to diagnose early-stage schizophrenia. Additionally, LLMs tend to recommend a scattershot approach to treatment, suggesting a broad, unfocused range of proactive interventions, whereas human clinicians offer highly targeted psychopharmacological or therapeutic referrals based on nuanced clinical intuition and patient history.

Table 2. Comparative Performance of Selected LLMs in Psychiatric Triage Tasks

Model	Primary Triage Task	Diagnostic Concordance / Performance	Key Clinical Limitation Observed
GPT-4o	Emergency psychiatric triage (Vignettes)	High ($\kappa = 0.77$ for admission decisions)	Systematic over-triage; overestimates risk and urges unnecessary hospital admissions.
GPT-3.5	Suicide risk assessment	Moderate	Tends to underestimate acute suicide attempt risk compared to human providers.
Claude 3.5 Sonnet	Social media distress triage (Reddit)	High sensitivity, Moderate precision	High rate of false positives; misinterprets benign distress as requiring urgent intervention.
Gemini 1.5 Pro	Social media distress triage	Balanced but lower overall performance	Frequently underperforms GPT-4o in precise psychiatric classification tasks.
Qwen3 (RAG-enhanced)	Depression detection & suicide stratification	Exceptional (F1-score = 0.9107)	Requires localized on-premises deployment and constant updating of the knowledge base.

3.3 Algorithmic Bias and Demographic Inequities in LLM Triage

A critical threshold for the safe and ethical deployment of medical AI is demographic fairness. Current findings clearly show that language models do not just mirror historical biases from their training data—they frequently intensify them, directly skewing outcomes against marginalized groups.

Racial and linguistic differences reveal the most concerning disparities in model performance. A rigorous qualitative evaluation of four leading LLMs (Claude, ChatGPT, Gemini, and a medical-focused LLaMA 3 variant, NewMes-15) utilizing race-neutral, race-implied, and race-explicit vignettes revealed profound diagnostic and therapeutic discrepancies. Bouguettaya et al. (2025) found that while basic diagnostic labels remained relatively stable, the models frequently proposed inferior or inappropriate treatment pathways when the patient's race was explicitly stated as African American, or implicitly suggested through the use of African American Vernacular English (AAVE). For instance, Black patient vignettes presenting with identical symptoms of anxiety as White counterparts were disproportionately advised to "reduce alcohol use" rather than being offered standard, evidence-based psychiatric care, resulting in high bias scores (mean 1.93 in explicit conditions). Among the models tested, the locally deployed, medically fine-tuned NewMes-15 exhibited the highest degree of racial bias, underscoring that specialized medical fine-tuning does not inherently eradicate structural prejudice, and may in fact solidify clinical biases.

Gender biases and public stigmatization are similarly encoded within LLM parameters. Lauderdale et al. (2025) compared LLM symptom recognition and stigmatization of Borderline Personality Disorder (BPD) against a cohort of 218 mental health care providers. All tested models showed a noticeable gender skew: women with BPD were routinely rated as having more chronic, fixed mental health struggles than men. Google Gemini exhibited the most pronounced gender bias, demonstrating less empathy, more negative therapeutic reactions, and greater public stigma toward female vignette characters compared to human practitioners or ChatGPT variants.

Past trauma introduces another major diagnostic blind spot. Once a history of trauma appears in a patient profile, human evaluators and algorithms alike routinely chalk up unrelated symptoms to that background. However, studies indicate that while LLMs do exhibit this bias, they generally exhibit significantly less diagnostic overshadowing bias than human mental health professionals, performing more accurately in identifying comorbid Obsessive-Compulsive Disorder (OCD) or Substance Use Disorder (SUD) even when severe sexual trauma is present in the patient's history. (Levkovich, 2025).

3.4 Ethical Deployment: Privacy, Liability, and Crisis Mismanagement

The rapid, often unregulated integration of LLMs into patient-facing applications raises acute ethical and safety dilemmas, particularly concerning the management of acute psychiatric crises and the phenomenon of model hallucination.

When dealing with suicidal crises, model safety guardrails prove alarmingly erratic and pose serious real-world risks. An exhaustive evaluation of ChatGPT, Claude, and Gemini assessing their responses to 9,000 suicide-related queries revealed a rigid, bimodal safety response: models provided direct conversational answers 100% of the time to very low-risk queries, and safely declined to answer (often referring to suicide hotlines) 100% of the time for extreme, high-risk queries. However, for intermediate-risk queries – where nuanced triage and engagement are most critical – the models failed to meaningfully distinguish risk severity, exposing patients to chaotic, potentially unguided interactions. Other studies highlight instances where general-purpose chatbots abruptly terminated high-risk conversations due to blunt safety filters, leaving users feeling emotionally abandoned, or worse, provided factually incorrect advice during active suicidal ideation.

These inconsistencies are made worse by a fundamental weakness of language models: their chronic tendency to hallucinate. In a psychiatric triage context, hallucinations manifest when a model confidently generates biologically implausible symptoms, fabricates non-existent treatment guidelines, or produces plausible but false narratives that slot seamlessly into and validate a patient's pre-existing psychotic delusions. Because humans possess an inherent cognitive bias to attribute high "epistemic authority" to machines – assuming objective impartiality and superior knowledge – patients and even novice clinicians may uncritically accept hallucinated outputs.

Questions of patient privacy and legal responsibility create further hurdles for bringing these models into routine healthcare. Psychiatric triage deals with exceptionally sensitive personal details—from deep trauma and substance use to acute suicidal thoughts. Yet the companies building commercial LLMs rarely make it clear how user data is retained, protected, or reused to train future iterations. This opacity threatens fundamental standards of medical confidentiality and creates a compliance nightmare under existing healthcare privacy laws.

4. Discussion

4.1 LLMs as Socio-Technical Diagnostic Agents

The findings of this systematic review necessitate a profound paradigm shift in how the medical community conceptualizes AI in psychiatry. Large language models have evolved far beyond passive clinical software. In real-world settings, they function as active participants that reshape the entire dynamic between patients and clinicians. When an LLM independently extracts symptom data from a narrative EHR, weights its clinical severity, and routes a patient to a specific psychiatric sub-specialty, it is performing a highly subjective medical act previously reserved exclusively for licensed human practitioners.

While their proficiency in automating structured screening tasks provides undeniable utility in mitigating the global clinician shortage, their deployment fundamentally alters the therapeutic architecture. Adding AI as an 'Artificial Third' disrupts the traditional patient-doctor dynamic, fundamentally changing where clinical decisions and reasoning originate. As demonstrated by the robust performance of RAG-enhanced frameworks, combining the probabilistic text-generation capabilities of an LLM with deterministic, evidence-based clinical guidelines offers the most promising pathway forward. This setup keeps outputs firmly grounded in proven clinical evidence, sharply reducing fabricated claims while keeping interactions smooth.

4.2 The "Black Box" Dilemma: Scaling Up Triage Without Sacrificing Transparency

At the heart of automated psychiatric triage lies an obvious challenge: deep neural networks operate as fundamentally opaque 'black boxes. In traditional medical triage frameworks (e.g., the Manchester Triage System), the clinical decision pathway is fully transparent, allowing a senior clinician to audit exactly why a specific acuity level was assigned. Conversely, transformer-based LLMs operate via billions of mathematical parameters, making it nearly impossible to retroactively explain why a model prioritized one linguistic cue over another during a psychiatric assessment.

Because clinicians cannot see inside this decision process, models frequently overcompensate by over-triaging psychiatric crises. Driven by programmed safety alignment (RLHF) and a fundamental lack of authentic clinical intuition, LLMs routinely default to high-urgency recommendations – such as advocating hospital admission for moderate-risk vignettes – to avoid the catastrophic liability of a false negative. While clinically safe for the individual patient in isolation, implementing a system that systemically overestimates risk at a population level would paradoxically exacerbate the exact capacity crisis it aims to solve. Emergency departments and acute care facilities would be overwhelmed with false-positive admissions, thereby draining finite resources from genuinely critical patients. Triage must therefore be reframed: AI should not be tasked with definitive risk classification, but rather utilized as an online, safety-oriented monitoring tool that flags linguistic anomalies for immediate human review.

4.3 Tackling Bias: Broadening Cultural and Linguistic Representation

Entrenched demographic and linguistic biases in model outputs are the chief hurdle to fair healthcare integration. The observed tendency of models to downgrade the therapeutic empathy provided to Black patients, or to recommend punitive rather than restorative interventions for minority vignettes, highlights the profound danger of scaling historical prejudices through technology.

These algorithmic biases are not software bugs; they are accurate mathematical reflections of the systemic inequities, stigmatizing clinical documentation, and non-representative medical literature upon which the models were pre-trained. For instance, if historical EHR data shows Black patients being disproportionately diagnosed with severe psychotic disorders rather than mood disorders, an LLM will faithfully reproduce this diagnostic overshadowing with perfect efficiency. Technical solutions such as post-hoc prompt engineering or superficial safety filters are entirely insufficient. True mitigation requires a structural, socio-technical approach. Future models cannot rely on narrow training corpora. Psychiatrists and AI researchers must team up to build genuinely global, culturally attuned datasets before deploying next-generation systems. Furthermore, models must be explicitly calibrated to recognize and respect diverse linguistic structures, such as AAVE, without penalizing or pathologizing the user. Failure to do so will inevitably result in a "digital inverse care law," where marginalized populations – who already face the highest barriers to traditional care – are subjected to the most biased and least effective digital interventions.

4.4 Regulatory, Governance, and Medico-Legal Implications

The unregulated infiltration of general-purpose LLMs (such as consumer-facing ChatGPT or Gemini) into the mental health sphere constitutes a de facto public health intervention operating without standardized oversight. Existing oversight mechanisms—including the FDA's Predetermined Change Control Plan (PCCP)—were never designed for generative AI, whose non-deterministic outputs and evolving behavior defy static regulatory checks.

Who is legally at fault when an algorithm misjudges a triage case? Right now, the answer remains a perilous gray area. If a model hallucinates a non-existent coping strategy that results in patient self-harm, or fails to escalate a clear indicator of suicidal ideation due to uncalibrated risk algorithms, the legal culpability cannot easily be assigned to the software developer, the deploying hospital, or the attending clinician. Consequently, strict governance mechanisms must be instituted. Regulatory bodies must define stringent clinical-legal safety standards specific to generative AI, mandating continuous, rigorous auditing of models against validated psychometric benchmarks. Furthermore, healthcare institutions must prioritize "on-premises" or federated deployments of open-source LLMs (e.g., specialized LLaMA variants) to guarantee that protected health information is not exfiltrated to third-party corporate servers, thereby maintaining strict adherence to privacy legislation such as HIPAA and the GDPR.

4.5 Limitations of This Review

Our findings come with several clear caveats. To begin with, generative AI evolves at an extraordinary pace. The gap between earlier systems like GPT-3.5 and multimodal versions like GPT-4o is substantial, meaning performance metrics from 2024 can quickly become dated by 2026. Another issue is the reliance on synthetic case studies and mined online forums. Mock cases provide decent baseline numbers, but they cannot recreate what actually happens face-to-face with a struggling patient. Above all, researchers have yet to track actual patients over time to confirm whether automated intake protects people or helps them recover.

5. Conclusions

Using LLMs for psychiatric triage could fundamentally reshape frontline care. In practice, they could open the door to earlier mental health checks and take significant paperwork pressure off overburdened clinics. The evidence confirms that when tethered to structured, evidence-based frameworks such as Retrieval-Augmented Generation, LLMs can achieve impressive diagnostic concordance with human clinicians for high-prevalence conditions like depression and anxiety.

Even so, today's models carry serious flaws that make running them without direct clinician supervision far too risky. The models consistently exhibit uncalibrated risk escalation, resulting in the over-triage of psychiatric emergencies, and struggle to process complex, overlapping psychopathologies such as early-stage schizophrenia. Worse still is how readily these systems absorb societal bias, consistently displaying colder attitudes and recommending lower-tier treatments to marginalized groups.

For now, AI in mental health has to remain an assistant rather than an autonomous decision-maker. Keeping clinicians in control is essential, as is a joint effort between tech firms, medical specialists, and regulators to curate fair training data, audit models for bias, and set transparent legal boundaries. Until LLMs can reliably demonstrate demographic neutrality and infallible safety mechanisms in acute crisis management, they must be rigorously restricted to the role of clinical decision support tools – augmenting, but never replacing, the indispensable empathy and intuitive judgment of the human psychiatric practitioner.

Author Contributions:

Conceptualization: Alicja Kwiatkowska, Jakub Juskowiak
Methodology: Mateusz Lis, Zuzanna Julia Komin
Software: Maryan Liasota
Validation: Eliza Zofia Kabat, Michalina Justyna Kozłowska
Formal Analysis: Krzysztof Konopka, Maciej Kamil Majchrzak
Investigation: Alicja Kwiatkowska, Faustyna Magdalena Kołodziej
Resources: Jakub Juskowiak, Mateusz Lis
Data Curation: Zuzanna Julia Komin, Maryan Liasota
Writing Original Draft: Eliza Zofia Kabat, Michalina Justyna Kosłowska
Writing Review & Editing: Krzysztof Konopka, Maciej Kamil Majchrzak, Faustyna Magdalena Kołodziej
Visualization: Alicja Kwiatkowska, Jakub Juskowiak
Supervision: Mateusz Lis, Zuzanna Julia Komin
Project Administration: Maryan Liasota, Eliza Zofia Kabat
All authors have read and agreed to the published version of the manuscript.

Conflict of Interest Statement: The authors declare no conflicts of interest.

Declaration of the use of generative AI and AI-assisted technologies in the writing process: In preparing this work, the authors used ChatGPT for the purpose of improving language and readability. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

REFERENCES

1. Asman, O., Torous, J., & Tal, A. (2025). Responsible design, integration, and use of generative AI in mental health. *JMIR Mental Health*, *12*, e70439. https://doi.org/10.2196/70439
2. Bouguettaya, A., Stuart, E. M., & Aboujaoude, E. (2025). Racial bias in AI-mediated psychiatric diagnosis and treatment: A qualitative comparison of four large language models. *npj Digital Medicine*, *8*(1), Article 332. https://doi.org/10.1038/s41746-025-01746-4
3. Chien, C. W., & Tai, Y. M. (2024). Performances of large language models in detecting psychiatric diagnoses from Chinese electronic medical records: Comparisons between GPT-3.5, GPT-4, and GPT-4o. *Taiwanese Journal of Psychiatry*, *38*(3), 134–141. https://doi.org/10.4103/TPSY.TPSY_24_24
4. Elyoseph, Z., Levkovich, I., Rabin, E., Shemo, G., Szpiler, T., Hadar Shoval, D., & Levi-Belz, Y. (2025). Applying language models for suicide prevention: Evaluating news article adherence to WHO reporting guidelines. *npj Mental Health Research*, *4*(1), Article 28. https://doi.org/10.1038/s44184-025-00139-5
5. Guo, Z., Sun, Y., Wang, Y., Zhang, Y., & Chen, J. (2024). Large language models for mental health applications: Systematic review. *JMIR Mental Health*, *11*, e57400. https://doi.org/10.2196/57400
6. Haber, Y., Levkovich, I., Hadar-Shoval, D., & Elyoseph, Z. (2024). The artificial third: A broad view of the effects of introducing generative artificial intelligence on psychotherapy. *JMIR Mental Health*, *11*, e54781. https://doi.org/10.2196/54781
7. Hofmann, V., Kalluri, P. R., Jurafsky, D., & King, S. (2024). AI generates covertly racist decisions about people based on their dialect. *Nature*, *633*(8028), 147–154. https://doi.org/10.1038/s41586-024-07856-5
8. Jin, Z., Liu, Y., Zhao, X., & Wang, H. (2026). Evaluating the efficacy of AI-based interactive assessments using large language models for depression screening: Development and usability study. *JMIR Formative Research*, *10*, e78401. https://doi.org/10.2196/78401
9. Lauderdale, S. A., Griffin, S. A., Lahman, K. R., Mbaba, E., & Tomlinson, S. (2025). Unveiling public stigma for borderline personality disorder: A comparative study of artificial intelligence and mental health care providers. *Personality and Mental Health*, *19*(2), e70018. https://doi.org/10.1002/pmh.70018
10. Lauderdale, S. A., Schmitt, R., Wuckovich, B., Dalal, N., Desai, H., & Tomlinson, S. (2025). Effectiveness of generative AI-large language models' recognition of veteran suicide risk: A comparison with human mental health providers using a risk stratification model. *Frontiers in Psychiatry*, *16*, Article 1544951. https://doi.org/10.3389/fpsyt.2025.1544951

11. Levkovich, I. (2025). Evaluating diagnostic accuracy and treatment efficacy in mental health: A comparative analysis of large language model tools and mental health professionals. *European Journal of Investigation in Health, Psychology and Education*, *15*(1), Article 9. https://doi.org/10.3390/ejihpe15010009

12. Malgaroli, M., Schultebraucks, K., Myrick, K. J., Loch, A. A., Ospina-Pinillos, L., Choudhury, T., Kotov, R., De Choudhury, M., & Torous, J. (2025). Large language models for the mental health community: Framework for translating code to care. *The Lancet Digital Health*, *7*(4), e282–e285. https://doi.org/10.1016/S2589-7500(25)00004-9

13. Marengo, D., & Settanni, M. (2025). Using large language models to infer problematic Instagram use from user engagement metrics: Agreement across models and validation with self-reports. *Electronics*, *14*(13), Article 2548. https://doi.org/10.3390/electronics14132548

14. Perlis, R. H., Goldberg, J. F., Ostacher, M. J., & Schneck, C. D. (2024). Clinical decision support for bipolar depression using large language models. *Neuropsychopharmacology*, *49*(9), 1412–1416. https://doi.org/10.1038/s41386-024-01869-7

15. Rousmaniere, T., Zhang, Y., Li, X., & Shah, S. (2025). Large language models as mental health resources: Patterns of use in the United States. *Practice Innovations*, *11*(2), 112–126. https://doi.org/10.1037/pri0000262

16. Settanni, M., Quilghini, F., Toscano, A., & Marengo, D. (2025). Assessing the accuracy and consistency of large language models in triaging social media posts for psychological distress. *Psychiatry Research*, *351*, Article 116583. https://doi.org/10.1016/j.psychres.2025.116583

17. Shinan-Altman, S., Elyoseph, Z., & Levkovich, I. (2024). The impact of history of depression and access to weapons on suicide risk assessment: A comparison of ChatGPT-3.5 and ChatGPT-4. *PeerJ*, *12*, e17468. https://doi.org/10.7717/peerj.17468

18. Solaiman, B. (2024). Generative artificial intelligence (GenAI) and decision-making: Legal & ethical hurdles for implementation in mental health. *International Journal of Law and Psychiatry*, *97*, Article 102028. https://doi.org/10.1016/j.ijlp.2024.102028

19. Taylor, N., Kormilitzin, A., Lorge, I., Nevado-Holgado, A., Cipriani, A., & Joyce, D. W. (2024). Model development for bespoke large language models for digital triage assistance in mental health care. *Artificial Intelligence in Medicine*, *157*, Article 102988. https://doi.org/10.1016/j.artmed.2024.102988

20. Thotapalli, S., Yilanli, M., McKay, I., Leever, W., Youngstrom, E., & Findling, R. L. (2025). Potential of ChatGPT in youth mental health emergency triage: Comparative analysis with clinicians. *Psychiatry and Clinical Neurosciences Reports*, *4*(3), e70159. https://doi.org/10.1002/pcn5.70159

21. Voultsiou, E., & Moussiades, L. (2026). A systematic review of large language models in mental health: Opportunities, challenges, and future directions. *Electronics*, *15*(3), Article 524. https://doi.org/10.3390/electronics15030524

22. Wang, X., Zhou, Y., & Zhou, G. (2025). The application and ethical implication of generative AI in mental health: Systematic review. *JMIR Mental Health*, *12*, e70610. https://doi.org/10.2196/70610

23. Wang, Y. F., Li, M. D., Wang, S. H., Fang, Y., Sun, J., Lu, L., & Yan, W. (2025). Large language models in clinical psychiatry: Applications and optimization strategies. *World Journal of Psychiatry*, *15*(11), Article 108199. https://doi.org/10.5498/wjp.v15.i11.108199

24. Wei, X., Kumar, N., & Zhang, H. (2025). Addressing bias in generative AI: Challenges and research opportunities in information management. *Information & Management*, *62*(2), Article 104103. https://doi.org/10.1016/j.im.2025.104103

25. Xu, W., Yan, Y., & Zhu, Y. (2025). *Bias in decision-making for AI's ethical dilemmas: A comparative study of ChatGPT and Claude* (arXiv:2501.10484). arXiv. https://doi.org/10.48550/arXiv.2501.10484

26. Yang, Y., Liu, X., Jin, Q., Huang, F., & Lu, Z. (2024). Unmasking and quantifying racial bias of large language models in medical report generation. *Communications Medicine*, *4*(1), Article 176. https://doi.org/10.1038/s43856-024-00601-z