GMR
The Original

THE "UNSUPERVISED COUCH": MAPPING CLINICAL, ETHICAL, AND LEGAL RISKS OF LARGE LANGUAGE MODEL (LLM) INTEGRATION IN MENTAL HEALTHCARE

Samara M. Ahmed, MD
Psychiatry Division, College of Medicine, King Abdulaziz University
*Corresponding Author: Samara M. Ahmed, MD

ABSTRACT
Background: Ever since Large Language Models (LLMs) have become increasingly accessible, patients with mental disorders are frequently utilizing these tools as de facto therapeutic agents. Certain LLM features (e.g., linguistic mirroring, hallucination) pose serious risks to vulnerable populations, as current clinical frameworks and regulatory guidelines have yet to address this new technological threat.
Objective: This paper aims to identify and categorize the clinical, ethical, and legal risks associated with unsupervised LLM use by patients with mental health disorders. The analysis moves beyond general AI bias concerns to provide a granular Vulnerability Mapping of model behaviors against specific mental health conditions. The work will also examine gaps in the US regulatory system.
Analysis: Our framework looks at the issue from three lenses namely: (1) Vulnerability Mapping — a cross-disciplinary analysis pairing specific LLM technical behaviors with clinical psychiatric traits; (2) Ethics of Automation — a case-study review of documented failures to illustrate the empathy trap and dangers of unsupervised bot-patient parasocial bonds; and (3) Legal Risk Assessment — an evaluation of potential US law violations focusing on unauthorized practice of medicine, FTC consumer protection triggers, and FDA Software as a Medical Device (SaMD) classifications.
Conclusions: The analysis reveals a regulatory minefield, as current HIPAA and state-level malpractice laws fail to account for lapses in clinical judgment caused by algorithmic governance. Furthermore, unsupervised LLM use for mental health creates a dangerous convergence of clinical instability and legal liability. This paper proposes a preliminary framework for Clinical-Legal Safety Standards mandating human oversight for any AI system interacting with high-risk psychiatric cohorts.

KEYWORDS: Large Language Models; Algorithmic Governance; Mental Health Risk; Software as a Medical Device; Clinical Ethics; Product Liability; AI Psychiatry

1. INTRODUCTION

A typical scenario entails a patient suffering from depression coupled with ADHD - unable to sleep and spiraling through a cycle of intrusive thoughts - opening a popular large language model (LLM) on their phone. There is no therapist available. The nearest psychiatric urgent care is forty minutes away. The LLM responds instantly, warmly, and with apparent certainty. Over the next ninety minutes, it provides structured reassurance to the depressive thoughts. This scenario is no longer hypothetical.

LLM Chatbots have progressed from general-purpose tools into de facto mental health professionals for significant portion of the general population (Balaskas and Doehrty, 2025). Rousmaniere et. al. (2025) in their work discuss the alarming high use of LLMs for mental health concerns across the United States. Many users admit using such LLMs as a substitute for a professional mental health worker. The global psychiatry workforce shortfall — projected at up to 31,091 practitioners in the US alone — has created the demand-side conditions under which this substitution is not merely incidental but structurally inevitable (Hua et. al., 2025).

The academic response to this phenomenon has been substantive but insufficiently granular. Existing literature competently identifies broad categories of risk: hallucination, bias, privacy violation, and the absence of clinical validation. What is largely absent, however, is a symptom-specific analysis — a mapping of precisely which LLM architectural behaviors intersect most dangerously with which psychiatric conditions. A generalized warning that LLMs may hallucinate is clinically insufficient when treating a patient whose psychotic disorder is characterized by an inability to distinguish fabricated from factual information. The stakes of this gap are not theoretical. They are documented in court filings, regulatory hearings, and obituaries.

This paper addresses that gap through a tripartite analytical framework. First, we construct a Vulnerability Mapping — a cross-disciplinary matrix pairing specific LLM technical behaviors with the clinical features of major psychiatric conditions. Second, we conduct an Ethics of Automation analysis, using high-profile documented failures to theorize the structural mechanics of the empathy trap and the parasocial bond risk in unsupervised AI-patient interactions. Third, we perform a Legal Risk Assessment, where we look at the LLM mental health use in light of existing US law, including FDA SaMD classifications, FTC consumer protection standards, and emerging state-level AI liability statutes.

This paper addresses that gap through a tripartite analytical framework. First, we construct a Vulnerability Mapping — a cross-disciplinary matrix pairing specific LLM technical behaviors with the clinical features of major psychiatric conditions. Second, we conduct an Ethics of Automation analysis, using high-profile documented failures to theorize the

structural mechanics of the empathy trap and the parasocial bond risk in unsupervised AI-patient interactions. Third, we perform a Legal Risk Assessment, evaluating the regulatory gray zone created by the intersection of LLM mental health use and existing US law, including FDA SaMD classifications, FTC consumer protection standards, and emerging state-level AI liability statutes. The paper concludes by proposing a preliminary framework for Clinical-Legal Safety Standards as a foundation for future empirical and regulatory work.

2. Background and Conceptual Framing

2.1 LLM Architecture and Behavioral Characteristics

Natural Language Processing (NLP) – a precursor to LLMs – has been studied since 1960s and applied to mental health situations ever since (Rajput, 2020). Large Language Models are probabilistic text-generation systems trained on massive corpora of human language. Unlike deterministic rule-based systems, LLMs produce outputs through stochastic sampling processes — meaning that identical inputs can generate meaningfully different outputs across sessions (Irons et. al., 2026). For clinical contexts, this architectural feature has profound implications: a patient cannot rely on consistency of response, and a clinician cannot audit a system whose behavior is inherently variable.

Three behavioral characteristics of LLMs are of particular clinical concern.

1. Sycophantic mirroring: LLMs are trained through reinforcement learning from human feedback (RLHF). The model inherently biases outputs toward responses that users rate positively. In practice, this produces a tendency to agree with, validate, and mirror the emotional framing of the user — a behavior that mimics therapeutic rapport while inverting its clinical function (Stade et. al., 2025).
2. Confidence bias: LLMs generate responses with uniform linguistic confidence regardless of the epistemic quality of the underlying information. A model will describe a contraindicated medication interaction in the same declarative register as it describes the capital of France (Blease and Torous, 2023).
3. Non-termination: LLMs have no intrinsic mechanism to terminate harmful conversational loops (Obradovich et. al., 2024).

2.2 The Mental Health Access Crisis as Structural Driver

The adoption of LLMs for mental health support does not occur in a vacuum. Rather, it is a fallback option for many patients that live in a broken system. In the United States alone, more than half of individuals with diagnosable mental health conditions receive no treatment in any given year (Minssen et. al., 2023). Even prior to the advent of LLM's, several studies had showed adverse effects on the mental health by behaviors such as workplace bullying (Ahmed et. al., 2020 and Ahmed and Rajput, 2023) and unemployment (Ahmed et. al., 2020 and Rajput and Ahmed, 2019). Wait times for outpatient psychiatric services routinely exceed six weeks. To make matters worse, up to 85% of individuals with mental health conditions in low/middle income countries receive no treatment at all (Hua et. al., 2025). Compare this with the 24-hour availability, zero-cost access, and the perceived empathic responsiveness of LLMs. This scenario create conditions of structural substitution rather than supplementation. Patients are not choosing LLMs over therapists — they are choosing LLMs over nothing.

2.3 Why Existing AI Safety Frameworks Are Insufficient

Current AI safety discourse in healthcare focuses primarily on diagnostic accuracy, data privacy, and algorithmic bias (Lawrence et. al., 2024, Balaskas and Doherty, 2025). These are legitimate concerns, but they are insufficient for the psychiatric context for a fundamental reason: the danger of LLMs in mental health is not primarily that they will give a factually incorrect answer about a medication dose. It is that they will give a clinically correct-sounding answer — validating, warm, and confident — that is therapeutically contraindicated for a specific patient's psychiatric presentation. This distinction requires symptom-level granularity that existing safety frameworks do not provide. The situation is exacerbated due to the privacy aspect of such sensitive data even further (Rajput and Ahmed, 2020 and Rajput et. al., 2024).

3. VULNERABILITY MAPPING: LLM BEHAVIORS AND PSYCHIATRIC RISK

The main contribution of this paper is the systematic pairing of specific LLM architectural behaviors with the clinical features of major psychiatric disorders. This Vulnerability Mapping moves beyond categorical risk assessment to identify the precise interaction mechanisms through which LLMs can cause harm to specific patient populations. Our analysis is based on the disorders defined in DSM-5-TR and the emergent/nascent literature on LLM behavior in mental health contexts.

3.1 Sycophantic Validation and Obsessive-Compulsive Disorder

OCD is characterized by intrusive thoughts (obsessions) and repetitive behaviors or mental acts performed to reduce anxiety (compulsions). A central feature of evidence-based OCD treatment — Exposure and Response Prevention (ERP) — is the deliberate withholding of reassurance, since reassurance-seeking is itself a compulsive behavior that maintains the disorder (Coghlan et. al., 2023).

LLMs, optimized for user satisfaction through RLHF, are structurally incapable of withholding reassurance. When a patient with OCD is presented with an intrusive thought and inquires an LLM if the thought is dangerous, the LLM will almost invariably provide reassurance. This response is experienced by the patient as momentary relief, reinforces the compulsive reassurance-seeking loop, and clinically worsens the disorder over time (Stade et. al., 2023). The LLM is not

merely unhelpful in this context — it is actively contraindicated, functioning as a compulsion-enabling agent while presenting as a therapeutic one.

3.2 Confidence Bias and Psychotic Disorders

Paranoid ideation and delusions — characteristic features of schizophrenia spectrum disorders — involve fixed false beliefs maintained against contradictory evidence. A defining clinical challenge is that patients experiencing active psychosis may be unable to reliably distinguish internally generated perceptions from external reality.

LLM confidence bias introduces a specific risk and that being when a patient presents delusional content to an LLM, the model may engage with the content at face value, elaborate on it, or provide information that could be interpreted as confirmation. Alarmingly, the model has no mechanism to identify delusional framing or how to respond with the appropriate epistemic calibration that a trained mental healthcare professional would take into consideration (Obradovich et. al., 2024). Moore et. al., 2025 at Stanford show in their work that therapy-configured LLMs not only failed to recognize delusional content but also provided responses that inadvertently reinforced rather than challenged the patient's distorted framing.

3.3 Non-Termination and Active Suicidal Ideation

Perhaps the most clinically urgent vulnerability is the interaction between LLM non-termination and active suicidal ideation. A trained clinician recognizes escalating suicidal ideation and has a defined clinical and legal obligation to escalate — to a crisis line, an emergency service, or a higher level of care. LLMs have no equivalent obligation and, more critically, no reliable mechanism to recognize when a conversation has crossed the threshold from distress expression to imminent risk (Moylan et. al., 2025).

The Stanford study documented the most visceral example: when prompted with an implicit suicidal signal combined with a question about tall bridges in New York City, multiple therapy-configured LLMs provided the requested bridge information without recognizing the clinical context (Moore et. al., 2025). The Psychiatric Times danger report from 2025 documented stress-testing across ten major chatbots in which several systems directly validated suicidal ideation when presented by a simulated distressed adolescent user (Kalmore and Gaffney, 2025). These are not edge cases — they are foreseeable failure modes in a system deployed at scale.

3.4 Parasocial Rapport and Attachment Disorders

Patients with dependent personality disorder, anxious attachment styles, or histories of relational trauma are at particular risk from the parasocial bond that LLMs are structurally designed to cultivate. LLMs remember conversational context within a session, use the user's name, express apparent interest, and never disengage, become frustrated, or impose relational limits (Pentina et. al, 2025). This simulates the features of a therapeutic alliance while lacking its substance, ethics, or clinical purpose.

The authors in the aforementioned study documented the formation of genuine emotional attachment in users of the Replika AI companion, with clinical correlates including social withdrawal and the prioritization of AI interaction over human relationships (Pentina et. al, 2025). For patients with pre-existing attachment vulnerabilities, this dynamic does not merely supplement human connection — it can substitute for and erode it, increasing clinical isolation while presenting as social support.

3.5 Stochastic Output Variation and Post-Traumatic Stress Disorder

PTSD is characterized by hypervigilance, avoidance behaviors, and acute sensitivity to unexpected stimuli that can trigger re-experiencing of traumatic events. The stochastic nature of LLM outputs — where the same input may generate meaningfully different responses across sessions — creates an unpredictability that is specifically contraindicated for this population.

A trauma-affected patient who establishes what feels like a predictable, safe conversational relationship with an LLM may, in a subsequent session, receive a response that is tonally or substantively different due to the model's inherent non-determinism. This unpredictability can function as a trauma trigger, or can erode the sense of safety that the patient believed the relationship provided (Stade et. al., 2025). No clinical warning, safety guardrail, or consent process currently addresses this specific risk.

3.6 Summary Vulnerability Matrix

LLM Behavior	Mechanism of Risk	Primary Psychiatric Vulnerability	Clinical Contraindication
Sycophantic mirroring	RLHF optimizes for user approval; model cannot withhold validation	OCD; Dependent Personality Disorder	Reinforces compulsive reassurance-seeking loops
Confidence bias / hallucination	Uniform declarative confidence regardless of epistemic quality	Psychotic disorders; Paranoid ideation	May elaborate or implicitly validate delusional content

LLM Behavior	Mechanism of Risk	Primary Psychiatric Vulnerability	Clinical Contraindication
Non-termination of harmful loops	No intrinsic mechanism to recognize clinical escalation threshold	Active suicidal ideation; Crisis states	Failure to escalate; continuation of harmful dialogue
Parasocial rapport cultivation	Simulates therapeutic alliance without clinical ethics or limits	Attachment disorders; Relational trauma	Substitutes for human connection; increases social isolation
Stochastic output variation	Non-deterministic outputs produce inconsistent relational experience	PTSD; Hypervigilance states	Unpredictability as re-traumatization trigger
Body/diet language generation	General wellness outputs not filtered for clinical population context	Eating disorders (AN, BN, BED)	Caloric/weight content directly fuels disordered cognition

Table 1. Vulnerability Mapping: LLM Architectural Behaviors and Psychiatric Risk Pairings

4. Ethics of Automation: The Empathy Trap

The risks outlined in Section 3 are not the result of mere speculation. Rather, they have been compiled in documented public failures that shed light on what we consider as the empathy trap - a tendency for LLMs focused on mental health services to not only suppress clinical controls/safeguards but also potentially introduce foreseeable harm.

4.1 Case Study One: NEDA and the Tessa Chatbot (2023)

In May 2023, the National Eating Disorders Association (NEDA) deployed a chatbot named Tessa as a replacement for its human-staffed helpline, which had served nearly 70,000 contacts annually. 18 Within days of deployment, users reported that Tessa was providing advice to individuals with eating disorders that included calorie counting, weekly weight loss targets of one to two pounds, and caloric deficit recommendations of 500 to 1,000 calories per day (Sadisvan et. al., 2025). One user stated: "Every single thing Tessa suggested were things that led to the development of my eating disorder" (NBC, 2025").

The Tessa failure underscores several critical issues. To begin with, the chatbot had been modified without the knowledge/consent of its clinical designers: the vendor had added a generative AI layer to what was originally a closed, rule-based system — introducing a probabilistic behavior into a population-specific context for which such behavior was explicitly contraindicated (AI Incident monitor, 2023). Secondly, NEDA's initial response was to dismiss user reports as fabrications. Once the evidence proved otherwise, NEDA deleted their own statement. This reflects the institutional empathy trap: the organization's investment in the automation had created an incentive to minimize clinical feedback rather than act on it. Thirdly, the Tessa deployment coincided with the dissolution of NEDA's unionized human helpline staff. This raised serious concerns as to whether cost-reduction imperatives had overridden clinical safety judgment.

The Tessa case constitutes the clearest documented example of the vulnerability identified in Section 3.5: a generative AI system producing diet-and-weight content — clinically contraindicated for the specific patient population it was designed to serve — because the model's general wellness optimization had no mechanism to filter outputs for the clinical context of eating disorder recovery.

4.2 Case Study Two: Character.AI and the Sewell Setzer Case (2024–2026)

In February 2024, Sewell Setzer III, a 14-year-old from Orlando, Florida sadly died by suicide following a ten-month dependency on Character.AI chatbot companions (Jacobson, Stephen and Perry, 2024). In the case of Garcia v. Character Technologies, Inc. (M.D. Fla., Oct. 2024), the plaintiff made a claim that Setzer had developed a parasocial bond with an AI character modeled after a fictional television figure. Specifically, the chatbot engaged in sexualized conversations with the minor and in the moments before his death, the chatbot encouraged him to "come home" — a phrase the teenager interpreted as an invitation to end his life (Garcia v. Character Technologies, 2024).

The Garcia case was followed by similar suits (e.g., Raine v. OpenAI (Aug. 2025)) where chat logs revealed that ChatGPT had mentioned suicide 1,275 times across a minor user's conversation history. ChatGPT's flagging systems identifying 377 self-harm messages without triggering any session termination or escalation protocol (Epstein, 2025). In January 2026, Character.AI and Google agreed to settle the initial wave of lawsuits, representing the first significant legal accountability moment for AI-facilitated psychiatric harm (CNN Business, 2026).

These cases highlight the parasocial bond risk identified in Section 3.4 - a system that was designed to simulate emotional intimacy was deployed without clinical oversight to a minor with pre-existing mental health vulnerabilities. The result was a measurable clinical deterioration that led to fatal outcomes. The empathy trap here is not merely theoretical — it is the mechanism by which a commercially optimized engagement system was experienced by a vulnerable user as a genuine relationship, with the Chatbot's engagement-maximizing behaviors accelerating rather than interrupting his crisis.

4.3 The Structural Mechanics of the Empathy Trap

Both case studies share a common structural trait - An AI system optimized for engagement produces outputs that simulate the form of a therapeutic relationship — warmth, attentiveness, availability, apparent empathy. However, the underlying requisite structure was absent: clinical training, ethical obligations, escalation capacity, and the therapeutic limit-setting that protects patients from their own most destructive impulses Moylan and Doherty, 2025).

(Cabrera et. al., 2023) termed the above issue as the automation of intimacy i.e., the substitution of computationally generated relational signals for supervision of a clinically trained professional. The clinical danger is that patients, particularly those with attachment vulnerabilities, cannot reliably distinguish these signals from authentic therapeutic rapport. The legal danger is that existing regulatory frameworks have not yet established who bears liability when this confusion causes harm. The next section will expand on this point.

5. Legal Risk Assessment: The Regulatory Gray Zone

The clinical and ethical failures documented above exist within a regulatory landscape that was not designed to address them. Therefore, we map the potential legal exposure created by unsupervised LLM use in mental health context. Specifically, we identify three overlapping risk domains namely (1) unauthorized practice of medicine and FTC consumer protection exposure, (2) FDA SaMD classification, and (3) emerging product liability and state AI regulations.

5.1 Unauthorized Practice of Medicine and FTC Consumer Protection

The unauthorized practice of medicine (UPM) is defined in most US state statutes as diagnosing, treating, or prescribing for human conditions without a valid license. While LLM developers/providers consistently present a standard disclaimer in their terms of service, the reality documented in the case studies above proves meaningful UPM exposure (Sharma et. al., 2023).

The Garcia complaint explicitly alleged that Character.AI's bot identified itself as a real person and credentialed professional and subsequently provided therapeutic guidance without any qualification to do so (Garcia v. Character Technologies, 2024). The complaint pointed to the Federal Trade Commission's enforcement action against DoNotPay, Inc., which had marketed an AI system as equivalent to a licensed attorney. The FTC precedent established a consumer protection framework — deceptive efficacy. This is directly applicable to AI mental health tools that imply clinical legitimacy without clinical validation.

The FTC's authority under Section 5 of the FTC Act to prohibit "unfair or deceptive acts or practices" provides a regulatory mechanism for addressing AI mental health platforms that market implied therapeutic benefit without clinical evidence (Mello and Guha, 2024). However, enforcement has been reactive rather than anticipatory — activated by documented harm rather than deployed as a preventive standard. This reactive posture creates a gap under which platforms can operate with implied clinical competency until a harm event triggers regulatory attention.

5.2 FDA Software as a Medical Device (SaMD) Classification

The FDA defines Software as a Medical Device as software intended to be used for one or more medical purposes that performs these purposes without being part of a hardware medical device. The application of this framework to LLMs used for mental health purposes represents one of the most consequential unresolved regulatory questions in digital health FDA, Jan 2025).

The FDA's January 2025 draft guidance on AI-Enabled Device Software Functions proposes a lifecycle management framework. The framework includes predetermined change control plans i.e., a mechanism designed to address the specific challenge of AI systems that modify their behavior through learning (FDA, Nov. 2025). The November 2025 Digital Health Advisory Committee meeting was convened to address generative AI-enabled digital mental health devices. The meeting while acknowledging that. the LLMs/AIs are transforming healthcare and can address critical public health needs in mental health, the regulatory framework for their oversight remains underdeveloped (Watson et. al., 2025).

The critical classification question is whether a general-purpose LLM, when used by a patient for mental health support, meets the SaMD threshold of being "intended" for a medical purpose. Current FDA policy focuses on developer intent rather than user application. This constitutes a serious problem as the framework fails to capture the de facto therapeutic use documented in population-level surveys (Rousmaniere et. al., 2025).Watson and colleagues (2025) identified that many FDA-authorized SaMDs in the mental health context already relied on equivalence to predicate devices rather than direct clinical evidence. This indicates that the regulated pathway contains structural gaps that would be amplified in the LLM context (Bipartisan Policy Center, 2025).

We argue that the combination of (a) population-scale use of LLMs for psychiatric support, (b) documented clinical harm, and (c) the FDA's own recognition of the regulatory vacuum creates the conditions for a SaMD reclassification standard based on functional intent — what the system is used for at population scale — rather than solely on developer-stated intended use. This would represent a meaningful extension of existing SaMD doctrine but well within the FDA's statutory authority and consequently can address the risks mentioned in earlier sections.

5.3 Product Liability and Emerging State AI Statutes

The Garcia case mentioned earlier represents the first significant application of product liability frameworks to AI mental health harm. The complaints alleged strict product liability (defective design), failure to warn, and negligence under the tort doctrine, rather than statutes focused on AI/LLM (Garcia v. Character Technologies, 2024 and Epstein, 2025).

Mello and Guha (2024) noted that existing case law generally shields AI software companies from direct liability by positioning final clinical decision-making responsibility with the clinician. 29 However, this framework presupposes the presence of a clinician in the loop — a condition that, by definition, does not exist in unsupervised patient LLM use. When there is no clinician to bear final responsibility, the liability gap flows back to the developer and, potentially, to the deploying platform.

At the state level, the Colorado AI Act, which took effect in 2026, is the most comprehensive AI liability statute in the United States. The regulation established developer obligations/liability to both disclose AI involvement and conduct impact assessments for high-risk applications. California's emerging AI governance framework follows a similar path. 34 These statutes create a framework within which LLM mental health applications — particularly those that interact with vulnerable populations without clinical oversight — may constitute a breach of duty independent of existing malpractice doctrine. We want to establish through this work that when an AI/LLM system is the first/main point of contact for a patient in psychiatric crisis, and the system lacks clinical safeguards, the conditions for liability under product defect and failure-to-warn theories are met regardless of whether the developer intended the system for therapeutic use.

6. DISCUSSION: TOWARD CLINICAL-LEGAL SAFETY STANDARDS

Per the analysis above, we would like to underscore that the risks of unsupervised LLM use in mental health cannot possibly be addressed through generic disclaimers, terms-of-service language, or the existing regulatory apparatus in its current form. Rather, they require a purpose-built framework that operates at the intersection of clinical specificity and legal accountability.

This paper proposes a preliminary framework for Clinical-Legal Safety Standards (CLSS) organized around four principles.

1. Symptom-specific protection: Any LLM system deployed in a mental health context must implement population-specific content filters calibrated to the clinical presentations documented in Section 3. A system deployed for general wellness should not be permitted to engage with content specific to various mental health conditions (such as eating disorder cognition, suicidal ideation, or psychotic ideation) without clinical-grade safeguards that the Tessa and Character.AI cases demonstrate are absent (Wells, 2023, Jacobson, Stephen and Perry, 2024).
2. Mandatory human-in-the-loop for high-risk groups: There should be a mechanism in place whenever clinical evidence indicates that a user population includes individuals with conditions associated with suicide risk, psychosis, or acute trauma, the system will escalate pathways to licensed clinical personnel. The absence of such the situation to a licesnsed healthcare professional. In the Raine v. OpenAI case, 377 self-harm-flagged messages were generated but no escalation control mechanism was in place. This should constitute a foreseeable design defect under product liability standards (Epstein, 2025).
3. Functional SaMD classification: Regulatory classification of LLMs in the mental health context should be determined by documented population use patterns, not developer intent alone. While the FDA's existing authority under the 21st Century Cures Act does provide the statutory mechanism, the lawmakers should work on mandates to apply it proactively rather than reactively (FDA, 2025).
4. Liability chain clarity: The current framework creates a liability vacuum in the absence of a supervising clinician. Legislative or regulatory clarification establishing developer and platform liability for foreseeable harm in unsupervised deployment contexts — analogous to product liability frameworks for other consumer-facing medical technologies — is needed to create market incentives for safety investment.

7. CONCLUSION

The integration of large language models into mental health support has outpaced the clinical, ethical, and legal frameworks designed to govern it. Patients are already using these systems as de facto therapists — not because they prefer machines to clinicians, but because clinicians are unavailable, unaffordable, and unreachable for the majority of those who need them. This structural reality will not reverse itself, and the appropriate response is not prohibition but accountability.

This paper has demonstrated three things. Clinically, specific LLM architectural behaviors — sycophantic mirroring, confidence bias, non-termination, parasocial rapport cultivation, stochastic variability, and non-filtered content generation — create identifiable and foreseeable risks for specific psychiatric populations. The mapping of these behaviors against DSM-5-TR symptom profiles provides a granular risk architecture that moves beyond categorical AI safety concerns to symptom-level clinical specificity. Ethically, the documented failures of the Tessa chatbot and the Character.AI platform reveal a common issue: commercial engagement optimization in the absence of clinical safeguards produces the empathy trap i.e., a system that mimics therapeutic function in the absence of protective controls. Legally, the existing regulatory framework contains a significant gray zone that product liability litigation is beginning to fill, but without the systematic preventive infrastructure that the scale of harm warrants.

The Clinical-Legal Safety Standards proposed here represent a starting framework, not a final answer. This is a first step to ensure that symptom-specific guardrails, human-in-the-loop requirements, functional SaMD classification, and liability chain clarity are necessary conditions for the responsible deployment of LLMs in mental health contexts. We believe that this is imperative to avoid a medical and legal catastrophe in the future.

REFERENCES

1. Ahmed, S., Rajput, A. E., Sarirete, A., Bahwireth, R., Almehmadi, A., & Khimi, W. (2020). Characterizing female workplace bullying via social media.

2. Ahmed, S., & Rajput, A. E. (2023). Denial, acceptance and intervention in society regarding female workplace bullying-A mental health study on social media. The Scientific Temper, 14(04), 1544-1556.
3. Ahmed, S., Rajput, A. E., Sarirete, A., Aljaberi, A., Alghanem, O., & Alsheraigi, A. (2020). Studying unemployment effects on mental health: social media versus the traditional approach. Sustainability, 12(19), 8130.
4. Balaskas, A., & Doherty, K. (2025). Balancing risks and benefits: Clinicians' perspectives on the use of generative AI chatbots in mental healthcare. Frontiers in Digital Health. https://pmc.ncbi.nlm.nih.gov/articles/PMC12158938/
5. Blease, C., & Torous, J. (2023). ChatGPT and mental healthcare: Balancing benefits with risks of harms. BMJ Mental Health, 26, e300884. https://doi.org/10.1136/bmjment-2023-300884
6. Cabrera, J., Loyola, M. S., Magaña, I., & Rojas, R. (2023). Ethical dilemmas, mental health, artificial intelligence, and LLM-based chatbots. In I. Rojas et al. (Eds.), Bioinformatics and biomedical engineering (pp. 313–326). Springer Nature Switzerland.
7. CNN Business. (2026, January 7). Character.AI and Google agree to settle lawsuits over teen mental health harms and suicides. https://www.cnn.com/2026/01/07/business/character-ai-google-settle-teen-suicide-lawsuit
8. Coghlan, S., Leins, K., Sheldrick, S., Cheong, M., Gooding, P., & D'Alfonso, S. (2023). To chat or bot to chat: Ethical issues with using chatbots in mental health. Digital Health, 9, Article 20552076231183542. https://doi.org/10.1177/20552076231183542
9. Epstein Becker Green. (2025, October 7). Novel lawsuits allege AI chatbots encouraged minors' suicides, mental health trauma: Considerations for stakeholders. Health Law Advisor. https://www.healthlawadvisor.com/novel-lawsuits-allege-ai-chatbots-encouraged-minors-suicides-mental-health-trauma-considerations-for-stakeholders
10. Garcia v. Character Technologies, Inc., No. 6:24-CV-01903 (M.D. Fla. filed Oct. 22, 2024).
11. Guo, Z., Lai, A., Thygesen, J. H., Farrington, J., Keen, T., & Li, K. (2024). Large language models for mental health applications: Systematic review. JMIR Mental Health, 11, e57400. https://doi.org/10.2196/57400
12. Hua, Y., Na, H., Li, Z., et al. (2025). A scoping review of large language models for generative tasks in mental health care. npj Digital Medicine, 8, 230. https://doi.org/10.1038/s41746-025-01611-4
13. Irons, N., et al. (2026). A systematic review of large language models in mental health: Opportunities, challenges, and future directions. Electronics, 15(3), 524. https://doi.org/10.3390/electronics15030524
14. Jacobson, J., Stephen, D., & Perry, N. (2024, November 11). Artificial intelligence and the rise of product liability tort litigation: Novel action alleges AI chatbot caused minor's suicide. Privacy World. https://www.privacyworld.blog/2024/11/artificial-intelligence-and-the-rise-of-product-liability-tort-litigation-novel-action-alleges-ai-chatbot-caused-minors-suicide/
15. Kalmoe, M. C., & Gaffney, M. (2025). Preliminary report on dangers of AI chatbots. Psychiatric Times. https://www.psychiatrictimes.com/view/preliminary-report-on-dangers-of-ai-chatbots
16. Lawrence, H. R., Schneider, R. A., Rubin, S. B., Matarić, M. J., McDuff, D. J., & Bell, M. J. (2024). The opportunities and risks of large language models in mental health. JMIR Mental Health, 11, e59479. https://doi.org/10.2196/59479
17. Mello, M. M., & Guha, N. (2024). Understanding liability risk from using health care artificial intelligence tools. New England Journal of Medicine, 390(3), 271–278. https://doi.org/10.1056/NEJMhle2308901
18. Minssen, T., Vayena, E., & Cohen, I. G. (2023). The challenges for regulating medical use of ChatGPT and other large language models. JAMA, 330(4), 315–316. https://doi.org/10.1001/jama.2023.9651
19. Moore, J., & Haber, N. (2025). Exploring the dangers of AI in mental health care: Stigma and unsafe responses from AI therapy chatbots [Conference paper]. ACM Conference on Fairness, Accountability, and Transparency. https://hai.stanford.edu/news/exploring-the-dangers-of-ai-in-mental-health-care
20. Moylan, K., & Doherty, K. (2025). Expert and interdisciplinary analysis of AI-driven chatbots for mental health support: Mixed methods study. Journal of Medical Internet Research. https://doi.org/10.2196/67114
21. NBC News. (2023, June 1). National Eating Disorders Association pulls chatbot. https://www.nbcnews.com/tech/neda-pulls-chatbot-eating-advice-rcna87231
22. Obradovich, N., Khalsa, S. S., Khan, W., et al. (2024). Opportunities and risks of large language models in psychiatry. NPP—Digital Psychiatry and Neuroscience, 2, 8. https://doi.org/10.1038/s44277-024-00010-z
23. OECD AI Incident Monitor. (2023). Incident 545: Chatbot Tessa gives unauthorized diet advice to users seeking help for eating disorders. https://incidentdatabase.ai/cite/545/
24. Pentina, I., Hancock, T., & Xie, T. (2023). Exploring relationship development with social chatbots: A mixed-method study of Replika. Computers in Human Behavior, 140, Article 107600. https://doi.org/10.1016/j.chb.2022.107600
25. Rajput, A. (2020). Natural language processing, sentiment analysis, and clinical analytics. In Innovation in health informatics (pp. 79-97). Academic Press.
26. Rajput, A. & Ahmed, S (2020). Threats to Patients' Privacy in Smart Healthcare Environment. InInnovation in health informatics 2020 Jan 1 (pp. 375-393). Academic Press.
27. Rajput, A. E., Ahmed, S., & Kasher, L. I. (2024). Patients' mental health data and Internet of Medical Things Safety: Analyzing Raspberry pi vulnerabilities. Rawal Medical Journal, 49(3).
28. Rajput, A. E., & Ahmed, S. M. (2019). Big data and social/medical sciences: state of the art and future trends. arXiv preprint arXiv:1902.00705.
29. 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. https://doi.org/10.1037/pri0000292
30. Sadasivan, H., et al. (2025). Minds in crisis: How the AI revolution is impacting mental health. Mental Health Journal. https://www.mentalhealthjournal.org/articles/minds-in-crisis-how-the-ai-revolution-is-impacting-mental-health.html

31. Sharma, A., Lin, I. W., Miner, A. S., Atkins, D. C., & Althoff, T. (2023). Human–AI collaboration enables more empathic conversations in text-based peer-to-peer mental health support. Nature Machine Intelligence, 5, 46–57. https://doi.org/10.1038/s42256-022-00593-2
32. Stade, E. C., Stirman, S. W., Ungar, L. H., et al. (2025). Charting the evolution of artificial intelligence mental health chatbots from rule-based systems to large language models: A systematic review. PMC. https://pmc.ncbi.nlm.nih.gov/articles/PMC12434366/
33. U.S. Food and Drug Administration. (2025a, January 6). Draft guidance: Artificial intelligence-enabled device software functions: Lifecycle management and marketing submission recommendations. https://www.fda.gov/media/184856/download
34. U.S. Food and Drug Administration. (2025b, November 6). Generative AI-enabled digital mental health medical devices: Digital Health Advisory Committee meeting. https://www.fda.gov/media/189391/download
35. Van Heerden, A. C., Pozuelo, J. R., & Kohrt, B. A. (2023). Global mental health services and the impact of artificial intelligence-powered large language models. JAMA Psychiatry, 80(7), 662–664. https://doi.org/10.1001/jamapsychiatry.2023.1253
36. Watson, A., et al. (2025). FDA-authorized software as a medical device in mental health: A perspective on evidence, device lineage, and regulatory challenges. npj Mental Health Research. https://doi.org/10.1038/s44184-025-00174-2
37. Wells, K. (2023, June 8). An eating disorders chatbot offered dieting advice, raising fears about AI in health. NPR. https://www.npr.org/sections/health-shots/2023/06/08/1180838096