# COVER PAGE AND DECLARATION

| | |
|---|---|
| | Master of Science (MS) in Mental Health Psychology |
| **Specialisation:** | Master Science |
| **Module Code & Module Title:** | MHP 700 Capstone Project |
| **Students' Full Name:** | Jamieson Dean Cobbs |
| **Word Count:** | 6084 |
| **Date of Submission:** | August 7, 2025 |

**I confirm that this assignment is my own work, is not copied from any other person's work (published/unpublished), and has not been previously submitted for assessment elsewhere.**

**E-SIGNATURE:** 

**DATE:** August 7, 2025

**EIU Paris City Campus**
**Address:** 59 Rue Lamarck, 75018 Paris, France | **Tel:** +33188320435 | **Mobile/WhatsApp:** +33607591197 | **Email:** paris@eiu.ac

**EIU Corporate Strategy & Operations Headquarter**
**Address:** 12th Fl. Amarin Tower, 496-502 Ploenchit Rd., Bangkok 10330, Thailand | **Tel:** +66(2)256923 & +66(2)2569908 | **Mobile/WhatsApp:** +33607591197 | **Email:** info@eiu.ac

# Master's Thesis

**Title:**

Friend, Therapist, or Threat? Exploring the Ethical and Clinical Implications of Large Language Models in Mental Health Practice

**Submitted by:**

Jamieson Cobbs

EIU6027075

**Program:**

Master of Science in Mental Health Psychology
European International University - Paris

**Date:**

August, 2025

Table of Contents

Abstract

Chapter 1: Introduction

1.1 Background of the Study

1.2 Problem Statement

1.3 Research Questions and Objectives

1.4 Justification for the Study

1.5 Significance of the Study

Chapter 2: Literature Review

2.1 Introduction

2.2 LLM Applications in Mental Health Practice

    2.2.1 Client-Led and Informal Uses

    2.2.2 Clinician-Led and Administrative Applications

    2.2.3 Gaps in Current Usage Data

2.3 Ethical Implications of LLMs in Mental Healthcare

    2.3.1 Privacy, Consent, and Data Protection

    2.3.2 Therapeutic Boundaries and Substitution Risk

    2.3.3 Risk of Misinformation and Hallucination

    2.3.4 Accountability and Professional Liability

2.4 Professional Guidelines and Ethical Standards Gaps

2.5 Conceptual and Practical Uncertainties

2.6 Summary and Conclusion

Chapter 3: Methodology

3.1 Research Design

3.2 Data Sources and Sampling Strategy

3.2.1 Document Types

3.2.2 Inclusion Criteria

3.2.3 Exclusion Criteria

3.3 Literature Search Strategy

3.4 Data Analysis: Thematic Synthesis

3.5 Researcher Reflexivity and Accountability

3.6 Limitations of the Methodology

3.7 Summary

Chapter 4: Data Analysis and Presentation of Results

4.1 Summary of Analysis Plan and Themes

4.2 Results for Objective 1: Current and Future Use of LLMs in Mental Health Practice

4.2.1 Client-Led Use: Digital Companionship and Self-Help

4.2.2 Clinician Use: Administrative Assistant and New Clinical Functions

4.2.3 Client and Clinician Use Summary

4.3 Outcome of Objective 2: Deficiencies in Ethical Issues and Professional Guidance

4.3.1 Inadequate Informed Consent and Privacy

4.3.2 Therapeutic Boundary Confusion and Anthropomorphism

4.3.3 Threat of Hallucination and Misinformation

4.3.4 Inability to Comply with Professional and Legal Standards

4.4 Result of Objective 3: Written Guidelines and Integration Models

4.4.1 Emerging Clinical Guidelines and Ethical Principles

4.4.2 Recommended Clinical Limits and Controls

4.4.3 Clinician-Centered Integration Models

4.4.4 Summary of Guideline Findings

Chapter 5: Conclusions and Recommendations

5.1 Summary of Findings

5.2 Implications for Mental Health Practice

5.3 Clinician-Focused Practice Recommendations

5.4 Future Research Directions

5.5 Conclusion

References

Abstract

The advent of large language models (LLMs) such as ChatGPT has opened up possibilities as well as raised ethical issues regarding mental health practice. This research explores clinical and ethical implications of LLM integration into mental health treatment through qualitative documentary analysis and thematic synthesis of peer-reviewed literature, ethical codes, policy reports, and professional guidelines between 2018 and 2025. This research aimed to (1) outline current and possible uses of LLMs by clinicians and clients, (2) evaluate ethical concerns of privacy, consent, disinformation, and therapeutic boundaries, and (3) draw pragmatic recommendations for ethical uptake.

Dominant themes included client-led digital companionship, clinician administrative use, consent ambiguity, therapeutic boundary erosion, misinformation risks, regulatory gaps, and integration frameworks. The conclusions suggest that LLMs are already changing the way individuals interact with and are provided with mental health care, often beyond professional monitoring or observable regulatory oversight. Clinicians cautiously use these technologies for non-clinical purposes, while ethical risks, such as hallucinated responses and data misuse, largely go unmitigated.

This research concludes that LLMs are promising but must be guided in their deployment. Recommendations include open transparency, human oversight, up-to-date informed consent, clinician AI literacy, and stakeholder-inclusive policy-making. Multiple case studies conclude that regulatory clarification is required and further empirical research in order to facilitate safe,

and ethical integration of LLMs into psychological practice.

# Chapter 1: Introduction

## 1.1 Background of the Study

The integration of artificial intelligence (AI) in mental health treatment is progressing at a fast pace with many benefits, but, equally as many ethical concerns. Large language models (LLMs) such as OpenAI's ChatGPT, in particular, have emerged as powerful apparatuses capable of interacting in natural language. Their instantaneous availability and conversational interface enable new forms of interaction ranging from casual aid and psychoeducation to more controversial uses like supplementation or substitution of therapy. Although the attraction of LLMs resides in their scalability and accessibility, their application in mental healthcare settings remains largely unregulated, crafting intricate clinical, ethical, and professional dilemmas.

LLMs have only been publicly available since late 2022, and their influence has become profound. Consumers are using them as surrogate companions, therapists are experimenting with their use in administrative and clinical tasks, and the public at large is increasingly turning to AI for emotional support. Despite this trend, there is little agreement in the literature regarding whether these instruments represent valuable adjuncts to care, dangerous substitutes for trained professionals, or entirely new entities requiring new ethical frameworks.

## 1.2 Problem Statement

As LLMs are increasingly used in mental health settings, there is a growing disparity between their rising use and professional consensus about their ethical, clinical, and practical implications. While some clinicians and clients feel that there are advantages to their use, others worry about misinformation, therapeutic relationship deterioration, privacy, and professional accountability. Professional regulatory agencies have offered no clear direction, and existing ethical guidelines were not authored with AI in mind. This creates a practice and literature vacuum that must be addressed.

## 1.3 Research Questions and Objectives

This research aims to:

1. Provide an overview of current and future applications of large language models (LLMs) in mental health settings, particularly in the context of clinician-client interactions with these technologies.
2. Assess the most suitable ethical concerns for clinical practice, including informed consent, therapeutic boundaries, and misinformation.
3. Integrate practice-based recommendations for how mental health clinicians can effectively and ethically implement LLMs in practice.

Based on these objectives, the research questions are:

What are the current and potential future uses of LLMs in mental health care?

What are the ethical issues and problems arising from these applications?

What guidelines can be gleaned from current literature to inform ethical and clinically sound use of LLMs in practice?

## 1.4 Justification for the Study

LLMs are transforming how people engage with mental health services, often without professional oversight. Clients may utilize them as therapeutic agents, and some clinicians are looking for ways to streamline practice or outsource tasks. But little is known about the psychological effect of these encounters, particularly in high-risk populations. Through careful analysis of this technology's clinical and ethical issues, this research will ascertain how mental health professionals can enter this terrain ethically. The rationale also exists in that literature is scattered, and often technologically focused, without mental health-specific consideration within established ethical frameworks.

## 1.5 Significance of the Study

This research adds to the increasing debate on AI in medicine with a mental health-specific, ethics-based analysis of LLM adoption. It meets an increasing need for guidance for mental health professionals, ethics boards, and policy-makers as they are increasingly confronted with treating clients calling on or using AI interventions in therapy. It also offers practical, peer-reviewed based recommendations and professional standards-based suggestions, which are currently lacking from discourse. The findings can also serve as a prelude to policymaking,

training efforts, and ethical revisions in a field rapidly evolving.

# Chapter 2: Literature Review

## 2.1 Introduction

The emergence of large language models such as ChatGPT, Claude, and Gemini has brought with it complex challenges as well as opportunities in mental health practice. We integrate existing literature on the use, impact, and ethical consideration of LLMs in mental health practice. The integration is themed to align with the three research objectives of the study and as a prelude to the thematic analysis in the proceeding chapters.

## 2.2 LLM Applications in Mental Health Practice

### 2.2.1 Client-Led and Informal Uses

Public access to LLMs has witnessed their utilization in emotionally supportive and self-help therapeutic communication. Dergaa et al. (2024) and Kasneci et al. (2023) note that clients are progressively utilizing LLMs as surrogate companions or therapists. Reddit communities such as r/mentalhealth and r/ChatGPT exhibit anecdotal patterns of clients using LLMs for emotional processing, anxiety alleviation, or even trauma revelation. While not clinically monitored, this use reflects a shift in how individuals engage with mental health services.

### 2.2.2 Clinician-Led and Administrative Applications

Clinicians convey a guarded view of the experimental use of ChatGPT and other large language models, but mainly for administrative purposes like the writing of session notes or the development of psychoeducational materials (Pratt, 2023). Although clinical trials are still few, early research and practitioner discussion indicate increasing interest in the use of artificial intelligence tools in independent practice, particularly for clients who need structured communication, including neurodivergent individuals. For instance, research into generative AI has examined its application in facilitating therapy planning and client engagement through the use of simplified language or script-building approaches (Wang et al., 2025). Nevertheless, systematic empirical assessments of clinical outcomes are still largely lacking. Issues of algorithmic fairness and the lack of transparency in decision-making processes have led scholars to call for the development of frameworks based on public reason and accountability (Binns, 2018).

#### 2.2.3 Gaps in Current Usage Data

There is no qualitative data as yet, and there are no longitudinal trials or studies validating the clinical effectiveness of LLMs for psychotherapy. The majority of research is either exploratory or user-perception studies. Academic interest remains mostly focused on usability versus therapeutic effect (Saeidnia et al., 2024)

### 2.3 Ethical Implications of LLMs in Mental Healthcare

#### 2.3.1 Privacy, Consent, and Data Protection

Client inputs to AI models can be stored or released to train future models. Informed consent,

data ownership, and adherence to privacy shields such as HIPAA or GDPR are some of the concerns this raises (Jobin et al., 2019; Sun et al. 2025). Anonymization has been reported by some LLM providers, yet technical research proves that metadata and prompt leakage risks are real.

### 2.3.2 Therapeutic Boundaries and Substitution Risk

LLMs can mimic therapeutic presence without empathy, responsibility, and human level of commitment. Anthropomorphism can be achieved with LLMs, with clients projecting therapeutic intention or emotional intelligence where there is none (Floridi & Chiriatti, 2020). Such boundary confusion endangers the human connection that is essential to therapy and can further exacerbate treatment avoidance in vulnerable patients.

### 2.3.3 Risk of Misinformation and Hallucination

LLMs are prone to factual hallucinations, highlighted by Bender et al. (2021). Their unreliability and neglect to prioritize truth have serious clinical implications, especially in high-stakes use cases like crisis response. Unlike trained practitioners, LLMs do not escalate or refer, and their suggestions are not grounded in evidence.

### 2.3.4 Accountability and Professional Liability

There is no clear model of liability when LLM-informed decisions result in injury. Would a practitioner be liable in the event that they had used ChatGPT to develop a plan of care with a negative outcome? What if the client injures himself after using AI advice? Professional

associations have yet to offer advice on this (EU AI Act, 2024).

### 2.4 Professional Guidelines and Ethical Standards Gaps

Professional ethical standards (APA, CPA, BPS) encompass broad competence and minimal harm principles but offer limited AI-specific guidance. None of them explicitly deal with LLMs in therapy as of 2025. There are some efforts: the APA 2023 Digital Ethics Working Group published a draft statement, and the EU AI Act (2024) offers general governance, but neither attains clinical specificity.

A comparative perspective by Tavory (2024) argues that while bioethical frameworks such as Beauchamp and Childress's Four Principles provide foundational guidance, they often fail to address the subtleties of LLM-specific issues, such as parasocial therapeutic relationships or AI-produced emotional influence, that impact mental health care. Furthermore, the translation of ethical principles into practical tools and guidelines has lagged, as observed by Morley et al. (2020), who highlight a gap between AI ethics theory and its implementation.

### 2.5 Conceptual and Practical Uncertainties

Clinicians are divided on the proper role of LLMs. Some view them as tools to expand access and streamline care, while others see them as threats to the therapeutic relationship and the profession itself. There is no consensus on their place in clinical practice. Marginalized communities are also underrepresented in AI datasets, elucidating concerns about bias and epistemic injustice (Whittaker et al., 2019).

Policy-level discussion also falls behind. National licensing boards, universities, and mental health training programs have not included AI use in codes, training curricula, or supervision models.

### 2.6 Summary and Conclusion

The literature shows a fast-evolving but fractured landscape. LLMs are already embedded in mental health discourse and behavior, yet professional infrastructures, empirical evaluations, and ethical policy fall behind. This review requires the need to develop coherent guidance according to ethical principles, informed by trends of emerging use, and with the objective of pragmatic integration for policy-makers, clients, and clinicians. The following chapters will synthesize this literature through thematic analysis to develop an ethically informed integration model of LLMs in mental health care.

## Chapter 3: Methodology

### 3.1 Research Design

This research employs qualitative documentary analysis to examine the deployment of large language models and their ethical evaluation in mental health settings. It was through this design that it was feasible to undertake systematic review and thematic synthesis of available academic, professional, and policy literature without primary data collection. The aim of this study is to

critically examine the role of LLMs in mental health practice and create ethically informed, practice-oriented recommendations. Documentary analysis is appropriate for this study because it allows for a systematic investigation of how LLMs are conceptualized, criticized, and controlled in tangible documents, such as peer-reviewed journal articles, codes of ethics, experts' comments, and policy statements. Thematic synthesis, as presented by Braun and Clarke (2006, 2021), was employed to identify recurring patterns and meanings across inter-disciplinary documents.

## 3.2 Data Sources and Sampling Strategy

### 3.2.1 Document Types

The following types of literature were included:

Peer-reviewed articles on AI ethics, LLMs, and mental health technology.

Professional ethics codes and guidelines (e.g., APA, CPA, BPS, ACA, EU AI Act).

Policy and governance reports by WHO, national AI ethics, and major mental health organizations.

White papers and position statement from research institute (e.g., AI Now Institute).

Peer-reviewed publications (e.g., Reddit r/mentalhealth, r/ChatGPT) only were published in peer-reviewed work or policy reports to portray evidenced user experience.

Theory and concept papers on bioethics, digital consent, and human-AI interaction.

### 3.2.2 Inclusion Criteria

Publications between 2018-2025 (to cover the transformer-era LLMs such as BERT, GPT-2, and ChatGPT).

English language.

Addressed at least one of the below mentioned: LLM mental health applications, ethical guidelines for AI, regulatory controversies that affect mental health AI, therapeutic implications, or clinical integration.

Very high preference for peer-reviewed sources; gray literature (white papers, policy reports) was included if they provided new or policy-relevant data.

### 3.2.3 Exclusion Criteria

Non-English materials.

News stories, non-formal blog entries, and unreferenced social media posts (unless addressed in peer-reviewed or policy reports).

Opinion pieces lacking empirical evidence, policy, or theory foundation.

Materials not relevant to mental health or not referring to LLMs or generative AI by name.

### 3.3 Literature Search Strategy

A systematic search was undertaken in the following academic databases and sources:

PsycINFO

PubMed

Scopus

IEEE Xplore (for technical ethics literature on AI)

Google Scholar (to locate gray literature, policy documents, and citation chaining)

Boolean search strings with keywords such as:

("large language models" OR LLM OR "generative AI" OR ChatGPT OR GPT OR BERT)

AND

("mental health" OR psychotherapy OR counseling OR "psychological practice")

AND

(ethics OR "informed consent" OR privacy OR accountability OR hallucination OR risk OR "AI regulation")

Searching was limited to 2018-2025 and English. Main journals (JMIR Mental Health, AI & Society, Ethics and Information Technology, etc.) were hand searched and backward/forward citation chaining used for coverage. Retrieved records were organized and de-duplicated in Zotero, and potential sources entered into a source matrix recording rationale for inclusion/exclusion.

### 3.4 Data Analysis: Thematic Synthesis

The Braun & Clarke six-phase method informed the analysis. Within the stage of familiarization, this consisted of thorough reading and note-taking regarding the given documents for immersion.

Initial Coding consisted of word-by-word and sentence-and paragraph-level identification of items in relation to clinical use, ethical concerns, professional recommendation, client outcome or policy recommendations. Descriptive labels such as "consent ambiguity", "AI administrative use," "hallucination risk" were utilized.

Theme Development: Codes were categorized into potential themes by research purposes (e.g., "boundary erosion," "data governance gaps").

To make them meaningful and distinctive, candidate themes were validated against the data. Peripheral or redundant themes were merged or remade.

After being finalized, themes were operationally defined and named in an effort to capture their analytic value. Finally, these themes were integrated into the finding's chapters (Chapters 4 and 5), aligned with evidence and practical guidance.

Coding and theming were followed inside a source-by-code matrix (spreadsheet) in a manner that allowed traceability and transparency.

### 3.5 Researcher Reflexivity and Accountability

Although without first-party participants, reflexive practice was employed in order to promote

analytic openness. During the period of analysis, the researcher employed a reflexive journal in which assumptions, interpretive choices, and ethical questions experienced were recorded, for instance, the dominance of Global North perspectives in the majority of the AI literature. Journaling was an internal, self-tracking system for tracking behind-the-scenes bias and organizing critical reading of sources.

In particular, reflexive journaling did not quite change the source pool (i.e., no additional searching or major change in included documents was caused by journaling as a direct effect). Instead, it affected interpretive sensitivity. It cautioned against overgeneralizing Western perspectives and identified gaps within the synthesis, such as the lack of Global South scholarship. This reflexive account preserves the rigor of the study in methodological limitation while demonstrating respect for ethical interpretation.

### 3.6 Limitations of the Methodology

The research is limited to policy and published literature with no collection of primary data and first-hand client or clinician feedback. Moreover, the field is evolving, and some policy or technical advance could have occurred subsequent to the cut-off date of 2025. Lastly, the literature remains biased toward Global North contexts, so generalizability of some results is limited.

### 3.7 Summary

The research strategy, systematic documentary review and Braun & Clarke-guided thematic synthesis, provided a systematic, open path to achieve the research purposes. Reflexive note-taking allowed interpretive rigorousness without overstating correction on the sample literature. The results are summarized in the subsequent chapter, arranged according to the purpose of this study and the themes extracted.

# Chapter 4: Data Analysis and Presentation of Results

## 4.1 Summary of Analysis Plan and Themes

This chapter presents the results of a systematic documentary review and thematic synthesis of the clinical and ethical implications of large language models for mental health practice. Analysis employed Braun and Clarke's six-phase approach to thematic analysis, which permitted inductive exploration of patterns across a diverse set of documents, including peer-reviewed articles, policy reports, ethical guidelines, and theory.

Thematic coding produced an organized matrix of recurring themes, arranged under the study's three research questions. Seven overarching themes were identified across all sources:

### Client-Led Utilization and Digital Companionship

### Clinician Use for Administrative and Therapeutic Work

Ambiguity of Consent and Privacy in AI Interactions

Blurring of Therapeutic Boundaries

Misinformation, Hallucination, and Safety Risks

Regulatory and Ethical Framework Gaps

Proposed Best Practices and Integration Models

These themes are arranged under the following sections:

Section 4.2 reviews how LLMs are currently being used (and may be used) by clinicians and clients.

Section 4.3 investigates the professional guidance gaps and ethical threats generated by such applications.

Section 4.4 reports on proposals for safe, ethical, and clinically relevant deployment of LLMs in mental health environments.

Direct quotes of primary sources are eschewed in favor of analytical description due to the character of data. However, references are given to ensure clarity and traceability of findings. Reflexive notes made while analyzing also led the synthesis, ensuring a sense of researcher assumptions, particularly on power, bias, and representation from the Global South.

## 4.2 Results for Objective 1: Current and Future Use of LLMs in Mental Health Practice

Primary analysis of literature reviewed identifies a nascent domain of LLM use in mental health, with clients and professionals using LLMs through informal and formal practice. Outcomes come under two underlying subthemes:

### 4.2.1 Client-Led Use: Digital Companionship and Self-Help

Primarily, users have begun to utilize LLMs, conversational interfaces like ChatGPT, to engage in emotionally supportive dialogue and self-referred mental health care. Users on platforms like Reddit's r/mentalhealth and r/ChatGPT refer to the use of LLMs as a "non-judgmental listener" or "therapist surrogate" when care is not available. They are informal and professionally unmonitored but typical of a behavioral mode in which LLMs are used as surrogates for friendship, emotional regulation, and discussion of personal adversity (Dergaa et al., 2024; Kasneci et al., 2023).

Some users described anthropomorphizing the model, stating, *"I tell ChatGPT things I'd never say to my therapist because it doesn't judge me."* Another shared, *"It feels like having a friend who's always awake when my thoughts spiral at night."* These anecdotal reports underscore the therapeutic illusion clients may form despite the lack of actual sentience or empathy in LLMs.

Although these are not therapeutic, many users appear to treat them as alive or capable of feeling empathy, suggesting growing parasocial relationship dynamics that are able to influence how individuals seek and receive mental health care. Some report feeling less isolated and emotionally relieved, but others are concerned about LLM errors or misinterpreting boundaries.

#### 4.2.2 Clinician Use: Administrative Assistant and New Clinical Functions

Clinicians have reported experimenting with LLMs for a range of non-clinical tasks such as session documentation, diagnostic code searches, and the generation of psychoeducational content. While rigorous clinical trials are still lacking, peer-reviewed evidence is emerging that highlights LLMs' potential to streamline administrative work and enhance accessibility, such as tailoring communication for neurodivergent or linguistically diverse clients (Wang et al., 2025). These applications remain largely exploratory, with ethical and effectiveness concerns still under investigation.

No controlled trials have yet determined safety or efficacy of the tools in treatment contexts. There are no large clinical trials, or peer-reviewed interventions, as of mid-2025. Most academic discussion hitherto has been exploratory or conjectural.

#### 4.2.3 Client and Clinician Use Summary

There is a bifurcated global trend in the literature: clients utilize LLMs directly for self-development, and clinicians increasingly start utilizing them as administrative aids more conservatively. All but one major evidence gap concerns the therapeutic impact of these models. Client utilization seen is prevalent but anecdotal, and clinician utilization is cautious and under researched. This initial ambivalence naturally segues into the ethical controversy of the second section.

### 4.3 Outcome of Objective 2: Deficiencies in Ethical Issues and Professional Guidance

The second objective of the study was to identify the general ethical issues related to the application of LLMs within mental healthcare settings. Four adequately represented thematic concern themes are set by the literature:

#### 4.3.1 Inadequate Informed Consent and Privacy

LLMs are accessed via interfaces that are not generally supplied with overt data control features. Their users are unaware that inputs may be stored, reused, or released during model training or model updating. Some of the ethical concerns are a lack of transparency on data ownership, unavailability of informed consent, and questionable compliance with data protection legislations like HIPAA, GDPR, or PIPEDA (Jobin et al., 2019; Sun et al., 2025; Blease & Rodman, 2025). Some platforms claim anonymization but technical studies confirm repeated breach such as instantaneous leak and term re-identification by metadata analysis.

This risk is compounded when LLMs are used in third-party apps or chatbots where backend data processing is non-transparent, once more undermining genuine informed consent.

#### 4.3.2 Therapeutic Boundary Confusion and Anthropomorphism

LLMs replicate empathic responses that are unemotional, not sentient, or ethically accountable. Vulnerable clients, in particular, might anthropomorphize such systems with intentionality or emotional nuance that is absent. The impact of this risk's client understanding of the therapeutic relationship and professional boundaries (Floridi & Chiriatti, 2020).

Boundary ambiguity is especially prevalent when users cite talking about sharing trauma, suicidal ideation, or abuse history with an AI, with some therapeutic presence assumed. This could discourage authentic help-seeking or cause over-reliance on unregulated sources.

#### 4.3.3 Threat of Hallucination and Misinformation

LLMs are fact-free. They produce coherent text but may produce factually incorrect, biased, or fabricated information, referred to as "hallucinations" (Bender et al., 2021). Large-scale tests in these dangers have documented specific harms ranging from stereotyping to disseminating risky health information (Weidinger et al., 2021), adding to the reasons why they are applicable to mental health safety.

In contrast to human counselors, LLMs are not executing duty-of-care responses such as crisis referral or reporting. Going unnoticed as artificial, their impact may create serious ethical

damage. This is reflected in recent work outlining that LLMs are not upholding the bare minimum of clinical standards of counseling, with hallucinations and wildly broad answers risking possible misdirection or patient confusion (Chung, Dyer, & Brocki, 2023).

#### 4.3.4 Inability to Comply with Professional and Legal Standards

Professional organizations (APA, CPA, ACA) have not issued binding standards for clinical use of LLMs. Ethical principles as a whole undeniably apply, e.g., competence, non-maleficence, and informed consent, but they were not written with generative AI in mind. Enforcement remains an issue, without an audit trail in private practice or penalizing non-compliance within clinics

Recent preliminary guidance by professional organizations and policymakers, such as initial digital ethics discussions in the APA and the EU AI Act (2024), is also too broad and non-clinically specific. Clinicians are unclear about liability, boundaries, and good practice in psychotherapy situations where there are no policies to apply and are specific to such situations.

Tavory (2024) suggests that while standard bioethics norms are not abandoned, they are lacking in the context of AI-facilitated emotional bonding, parasocial trust, and algorithmic manipulation in mental health contexts.

The EU AI Act (2024) categorizes AI systems used in mental health as “high-risk,” requiring explicit documentation, human oversight, and transparency in their use. Article 13 mandates that

users be informed when they are interacting with an AI and that decisions affecting human well-being not be left to systems themselves. Encouraging as this sounds, it's still a very general framework with no particular clinical enforcement measures in place, especially outside the EU. Its actual adoption among mental health practitioners, particularly in solo or small-group private practice, has yet to be quantified. Moreover, its jurisdictional extent suggests the need for global AI ethics harmonization among regulatory bodies.

### 4.4 Result of Objective 3: Written Guidelines and Integration Models

The study of objective 3 was to determine and synthesize better practice guidance for the clinically informed and ethical integration of LLMs in the treatment of mental health. The research identified a growing body of clinician-oriented guidance aimed at addressing the safe and ethical use of LLMs in therapeutic contexts. This guidance centers on practical strategies to uphold informed consent, maintain clear therapeutic boundaries, and safeguard clients during AI-assisted mental health care.

#### 4.4.1 Emerging Clinical Guidelines and Ethical Principles

Professional associations and clinical ethicists have begun to propose foundational guidelines for therapists who would like to integrate AI responsibly into practice. The models emphasize clinician education, client disclosure, and avoiding use of AI for unsupervised therapeutic decision-making. Professional association and clinical ethicist guidelines emerging recently emphasize clinician education about AI, client disclosure, and avoiding unsupervised therapeutic decision-making with AI technologies. Although early deliberations (such as those reportedly

informal in APA working groups) are said to reflect increasing recognition, formal, enforceable policy is still underway.

Similarly, the EU AI Act (2024) suggests risk classification of AI systems and encourages human control in such use where high-risk is present, e.g., mental health support.

These reports are the start of a governance framework but barely touch on providing fully actionable clinical recommendations.

Tavory (2024) and Whittaker et al. (2019) both highlight the necessity of possessing ethical accountability infrastructures, such as audit trails or data justice mechanisms, to balance structural bias and protect underrepresented groups in AI systems designed to be used in mental health.

### 4.4.2 Recommended Clinical Limits and Controls

The literature calls for clinical limits to be established when integrating LLMs. These are:

Transparency to Clients: Inaccurate reporting on the usage and limitations of AI in sessions or client materials.

Human Supervision: Having any material created by AI utilized in clinical practice overseen and corrected by a skilled practitioner.

Informed Consent Addenda: Developing consent forms discussing the application of AI, risk of data, and defines clinician responsibility.

There also exists the issue that AI systems must never be presented as a replacement for therapists. Even in administrative roles, LLM outputs need to be reviewed and clinicians always need to have ultimate decision-making authorities to uphold professional responsibility.

#### 4.4.3 Clinician-Centered Integration Models

There exist blueprints or integration roadmaps by some authors, e.g.:

Human-in-the-Loop Models: Passing through human clinicians AI recommendations, particularly for decision-making or patient communication functions.

Training and Competence Models: Suggesting certification or courses in AI literacy and digital ethics within clinical training.

Multistakeholder Involvement: Ensuring guideline development through contributions from technologists, mental health professionals, service users, and ethicists to facilitate varied perspectives and avoid epistemic bias.

While still emerging, these clinician-centered models offer a clear direction: ethical LLM integration must go beyond technical controls to include clinical judgment, practitioner accountability, and robust understanding of therapeutic boundaries. Therapists must retain decision-making authority and communicate AI use transparently to avoid eroding client trust.

#### 4.4.4 Summary of Guideline Findings

Though scarce, the literature here points toward more consensus for a recommendation that ethical integration of LLMs prioritize transparency, human oversight, informed consent revised to reflect today's conditions, and AI literacy by practitioners. Best practice is hypothetical at present, but foundations are being laid.

## Chapter 5: Conclusions and Recommendations

### 5.1 Summary of key findings

The author set out to examine the clinical and ethical matters regarding the use of large language models in mental healthcare through documentary review and thematic synthesis. Three basic aims informed the study: determining how LLMs are presently utilized in mental healthcare, investigating the ethical concerns in their usage, and synthesizing guidelines for best practice in their ethical use.

Seven overarching themes emerged:

Client-Led Use and Digital Companionship

Clinician Utilization for Administrative and Therapeutic Tasks

Unclearness of Consent and Privacy in AI Interactions

Therapeutic Boundary Blurring

Misinformation, Hallucination, and Safety Concerns

Gaps in Regulatory and Ethical Frameworks

Suggested Best Practices and Models of Integration

Cumulatively, these subjects paint a picture of quicker-than-anticipated uptake, ethical uncertainty, and a need for proactive professional direction.

## 5.2 Implications for Mental Health Practice

Consequently, the study focuses on the emphasis that LLMs have already revolutionized engaging the mental health setting, both formally and informally. Clients typically respond to the tools in an affectively responsive manner as humans would; these may cause parasocial relationships leading to modifying help-seeking behavior. Clinicians are more likely to adopt LLMs tentatively to administer the tools than integrate them in clinical decision-making.

Uncertainty of professional and ethical direction puts practitioners at risk of liability and their clients at risk of harm. Professional organizations have begun to respond, but their direction is vague and not specific to frontline practice.

To integrate LLMs into mental health practice ethically, clinicians and organizations need to co-create evidence-based protocols, prioritize clinician AI literacy, and apply ethical principles to changing situations.

## 5.3 Clinician-Focused Practice Recommendations

The following are provided as actionable recommendations corresponding to the literature reviewed:

Transparency and Disclosure: Clinicians should be candid in disclosing any use of AI tools, and then explain the restriction, for example, "This chatbot was developed with the help of AI. It is

not a replacement for human care."

Human Monitoring: LLM output should always be monitored and interpreted by a licensed professional. Example: "All AI-generated materials (e.g., psychoeducation, summaries) should be reviewed prior to use." This aligns with implementation science literature - Sendak et al. (2020) advocate for context-dependent adaptation and oversight of machine learning systems in healthcare.

Modified Informed Consent must now cover AI-specific risks.

e.g., "AI tools may process and store information; outputs are advisory and may not be evidence-based."

AI Competence: Clinicians should continue training in AI ethics and function to ensure safe integration.

Role Delineation: LLMs must never substitute for clinicians in therapeutic decision-making.

e.g., "AI-generated treatment suggestions should not be used without human clinical judgment."

Collaborative Protocol Design: Clinical teams need to collaborate on explicit internal policies for the use of LLMs, which will decrease liability and provide solidified ethical standards among

practitioners.

### 5.4 Future Research Directions

Given the fluidity of AI technologies, the following lines of research must be developed:

Empirical Research: Quantitative and qualitative research on the outcomes of clients resulting from LLM-supported therapy or emotional support.

Crisis Response Models: Research on whether and how LLMs might effectively exacerbate mental health crises in a safe and ethical manner.

Regulatory Impact Studies: Assessment of how policies such as the EU AI Act affect practice and liability in the clinical area.

Marginalized Populations: Examinations of the impact of LLMs on, or the marginalization of diverse populations in mental health care delivery.

### 5.5 Conclusion

LLMs came onto the scene of mental health with an audacious speed before ethical consensus and clinical support catch up. They promise, first and foremost, access and efficiency; but they

also pose rather basic ethical challenges touching on privacy, professionalism, misinformation, and human relationship. This study contributes a systematic analysis of these apprehensions and proposes a framework of emerging guidelines for ethical adoption. Successful forward momentum, therefore, will demand that mental health professionals approach this compelling but incomplete technology in critical, ethical, and collaborative ways.

References

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, *2021*, 610–623. https://doi.org/10.1145/3442188.3445922

Binns, R. (2018). Algorithmic accountability and public reason. *Philosophy & Technology*, *31*(4), 543–556. https://doi.org/10.1007/s13347-017-0263-5

Blease, C., & Rodman, A. (2025). Generative artificial intelligence in mental healthcare: An ethical evaluation. *Current Treatment Options in Psychiatry*, *12*(5). https://doi.org/10.1007/s40501-024-00340-x

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Braun, V., & Clarke, V. (2021). *Thematic analysis: A practical guide*. SAGE Publications.

Chung, N. C., Dyer, G., & Brocki, L. (2023, November 23). Challenges of large language models for mental health counseling. *arXiv*. https://doi.org/10.48550/arXiv.2311.13857

Dergaa, I., Fekih-Romdhane, F., Hallit, S., Loch, A. A., Glenn, J. M., Fessi, M. S., Ben Aissa, M., Souissi, N., Guelmami, N., Swed, S., El Omri, A., Bragazzi, N. L., & Ben Saad, H. (2024). ChatGPT is not ready yet for use in providing mental health assessment and interventions. *Frontiers in Psychiatry*, *14*, 1277756. https://doi.org/10.3389/fpsyt.2023.1277756

European Union. (2024). *EU Artificial Intelligence Act*. https://artificialintelligenceact.eu/

Floridi, L., & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines*, *30*(4), 681–694. https://doi.org/10.1007/s11023-020-09548-1

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*, 389–399. https://doi.org/10.1038/s42256-019-0088-2

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., . . . Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, *103*, 102274. https://doi.org/10.1016/j.lindif.2023.102274

Morley, J., Floridi, L., Kinsey, L., & Elhalal, A. (2020). From what to how: An initial review of publicly available AI ethics tools, methods and research to translate principles into practices. *Science and Engineering Ethics*, *26*(4), 2141–2168. https://doi.org/10.1007/s11948-019-00165-5

Pratt, N. (2023). ChatGPT: Artificial Intelligence & Mental Health. *Psychiatric Times*. https://www.psychiatrictimes.com/view/chatgpt-artificial-intelligence-mental-health

Saeidnia, H. R., Hashemi Fotami, S. G., Lund, B., & Ghiasi, N. (2024). Ethical Considerations in Artificial Intelligence Interventions for Mental Health and Well-Being: Ensuring Responsible Implementation and Impact. *Social Sciences, 13*(7), 381. https://doi.org/10.3390/socsci13070381

Sendak, M. P., D'Arcy, J., Kashyap, S., Gao, M., Nichols, M., Corey, K., & Ratliff, W. (2020). A path for translation of machine learning products into healthcare delivery. *EMJ Innovations.* https://doi.org/10.33590/emjinnov/19-00172

Sun, Q. W., Miller, J., & Hull, S. C. (2025). Charting the ethical landscape of generative AI augmented clinical documentation. *Journal of Medical Ethics*. https://doi.org/10.1136/jme-2024-110656

Tavory, T. (2024). Regulating AI in mental health: Ethics of care perspective. *JMIR Mental Health*, *11*, e58493. https://doi.org/10.2196/58493

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

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P. S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., Kenton, Z., Brown, S., Hawkins, W., Stepleton, T., Biles, C., Birhane, A., Haas, J., Rimell, L., Hendricks, L. A., … Gabriel, I. (2021, December 8). *Ethical and social risks of harm from language models*. arXiv. https://doi.org/10.48550/arXiv.2112.04359

Whittaker, M., Crawford, K., Dobbe, R., Fried, G., Kaziunas, L., Mathur, V., West, S. M., Richardson, R., Schultz, J., & Schwartz, O. (2019). *Disability, bias, and AI: Policy recommendations for fairness and justice*. AI Now Institute. https://ainowinstitute.org/wp-content/uploads/2023/04/AI_Now_2019_Report.pdf