The Role of Large Language Models in Mental Health: A Scoping Review

Tiago Gomes

Dissertation written under the supervision of

Professor Henrique Martins

Dissertation submitted in partial fulfilment of requirements for the MSc in Business, at the Universidade Católica Portuguesa, September 12rd 2024.

Table of Content

Abstract

Mental health disorders affect nearly one billion individuals worldwide, with a growing prevalence over year, caused in part due to stigma and lack of treatment causing a high burden for healthcare systems. In this context, Large Language Models (LLMs), such as GPT-4, have emerged as transformative tools with the potential to improve mental health care. This master thesis conducts a scoping review of research published from 2023 onwards to explore the current applications of LLMs within the realm of mental health, with the objective of offering a thorough overview of their existing and prospective applications in clinical practices and data analysis. While LLMs hold promise in improving mental healthcare through early diagnosis, treatment planning, and the communication between patients and clinicians, this review has also pointed out the limitations the current models have, such as the high-risk mental health crisis, an inability to understand emotional subtleties which are crucial in the treatment of mental health, and concerns about ethics and data privacy in relation to the inherent biases of the training data. For future research, key areas include enhancing LLMs' skills in recognizing crises, creating tailored models for mental health for higher sensibility, and addressing significant ethical issues like bias and data privacy, which are essential for the gradual integration into the mental health field. LLMs integration in the mental health sector require a careful integration in order ensure patient safety and maintaining trust. It is imperative to have human oversight while using these tools, especially in high-risk clinical environments.

Title:
O Papel dos Grandes Modelos Linguísticos na Saúde Mental: Revisão Escopo

Author: Tiago Gomes

Heywords: Large Language Models (LLMs); Mental Health; Applications; Clinical Data Analysis; Generative Pre-Training (GPT); Screening; Risk Detection; Treatment; Recommendations; Ethical Challenges; Data Privacy; Communication; Therapeutic Interventions; Natural Language Processing (NLP); Artificial Intelligence

Resumo

Perturbações de saúde mental afetam cerca de um bilião de pessoas mudialmente, com uma prevalência anual crescente, em parte devido ao estigma e insuficiência dos tratamentos, representando um elevado encargo para os sistemas de saúde. Neste contexto, os grandes modelos de linguagem (GMLs), como o GPT-4 surgiram como ferramentas inovadoras com potencial de melhorar os cuidados de saúde mental. Esta dissertação realiza uma revisão de escopo de estudos publicados a partir de 2023 que exploram as aplicações atuais de GMLs no domínio da saúde mental, com o objetivo oferecendo uma visão geral completa das suas aplicações atuais no contexto de práticas clínicas e análise de dados. Embora os GMLs demonstrem potencial para melhorar a saúde mental, em áreas como diagnóse prévia, planeamento clínico e promoção da comunicação paciente-médico, aponta também aponta algumas limitações, como a dificuldades em lidar com crises de alto risco, não compreender as subtilezas emocionais essenciais às conversas sobre saúde mental, questões éticas relacionadas com a privacidade dos dados e enviesamentos presentes nos dados de treino. Para investigação futura, as área principais incluem o reforço das competências dos GMLs no reconhecimento de crises, criação de modelos adaptados à saúde mental e abordagem de questões éticas, como o enviesamento e a privacidade dos dados, que são essenciais para a integração gradual no domínio da saúde mental. A integração dos GMLs neste setor exige uma integração cuidadosa, garantindo a segurança dos pacientes e a sua confiança. É imperativo que exista supervisão humana, especialmente em ambientes clínicos de alto risco.

Título: O Papel dos Grandes Modelos Linguísticos na Saúde Mental: Revisão Escopo

Autor: Tiago Gomes

Descritores: Grandes Modelos de Linguagem (GMLs); Saúde Mental; Aplicações; Análise de Dados Clínicos; Pré-treino Generativo (GPT); Rastreio; Deteção de Risco; Tratamento; Recomendações; Desafios Éticos; Privacidade dos Dados; Comunicação; Intervenções Terapêuticas; Processamento de Linguagem Natural (PLN); Inteligência Artificial

List of Abbreviations

LLM - Large Language Model

GPT - Generative Pre-Training

NLP - Natural Language Processing

AI - Artificial Intelligence

HIPAA - Health Insurance Portability and Accountability Act

GDPR - General Data Protection Regulation

CBT - Cognitive Behavioral Therapy

WHO - World Health Organization

PST - Problem-Solving Therapy

PHQ-9 - Patient Health Questionnaire-9

PRISMA - Preferred Reporting Items for Systematic Reviews and Meta-Analyses

NN - Neural Networks

LSTM - Long Short-Term Memory

BERT - Bidirectional Encoder Representations from Transformers

T5 - Text-To-Text Transfer Transformer

FLAN - Fine-tuned Language Net

ROUGE - Recall-Oriented Understudy for Gisting Evaluation

IoMT - Internet of Medical Things

List of Tables

List of Figures

List of Exhibits

1 Introduction

Mental health disorders, including anxiety, depression, and bipolar disorder, represent significant global health issues, affecting approximately one in eight individuals worldwide and leading to considerable economic and societal challenges. The implications of these conditions extend significantly, impacting both affected individuals and healthcare systems, as well as economies. The World Health Organisation reports that mental health disorders result in an annual economic loss exceeding $1 trillion, attributed to decreased productivity (WHO, 2019). While mental health is increasingly acknowledged as a significant public health concern, numerous challenges persist, especially in the areas of diagnosis, treatment, and patient engagement. Mental health care frequently depends on subjective, self-reported data and conventional assessment techniques, potentially leading to delays in diagnosis and treatment, especially in cases such as bipolar disorder, which may require years for accurate diagnosis.

Recent advancements in Artificial Intelligence (AI), particularly in Large Language Models (LLMs), have sparked considerable interest in how they might change the landscape of mental health care. Large language models like OpenAI9s GPT-4 and Meta9s LLaMA show great promise in a range of tasks, including logical reasoning and understanding language. The ability to create text that feels human, analyse large amounts of language data, and provide meaningful insights holds great promise for use in various fields, such as mental health. Bringing LLMs into mental health care is especially important given the difficulties the mental health field faces, like limited resources and the lack of access to qualified professionals. In resource-limited settings, large language models may function as effective instruments for early diagnosis, treatment planning, and therapeutic interventions, aiding clinicians in delivering personalised care to wider populations. This thesis examines the capacity of LLMs to enhance mental health applications, specifically within clinical environments and data analysis. Furthermore, it seeks to investigate the utilisation of LLMs for early detection of mental health conditions, diagnostic support, treatment recommendations, and improvement of patient-clinician communication. The existing literature on the application of LLMs across various domains is expanding; however, the area of mental health is still insufficiently examined. This thesis reviews studies that elucidate the opportunities and limitations of LLMs, especially regarding their capacity to integrate automated decision-making with the emotional and contextual nuances critical to mental health care.

It is important to note that the intricate and delicate aspects of mental health care necessitate that the implementation of LLMs considers critical ethical issues, such as data privacy, bias, and the potential dangers of excessive dependence on AI-driven models. Lastly, this thesis analyses recent research on the application of LLMs in mental health care, providing deeper insights and recommendations for future research and ethical considerations.

2 Background

2.1 Large Language Models

Large Language Models (LLMs), more specifically Generative Pre-Training models (GPT), have gained global attention for their ability to cognitive tasks like logical reasoning, text summarization, and language understanding (Baktash & Dawodi, 2023). In these models, the user 8talks9 to the LLM as if it was a virtual assistant, by describing the task they want to be performed. This capability opens the possibility of application across numerous fields, including mental health, where these models are being tested to improve the understanding and responsiveness to patient needs (Brown et al., 2020). Yann LeCun, an early adopter of this field, emphasizes that these models' strength is their ability to "learn representations" of the complex structure of human language and transform this raw data into structured and meaningful information. (LeCun, Bengio, & Hinton, 2015).

A number of extremely potent conversationally optimized LLMs, like Anthropic's Claude or OpenAI's ChatGPT, are accessible to the general public. However, the internal training data of the models are not publicly available classifying them as closed-source models. This lack of transparency raises concerns regarding user data privacy, copyright violations, regulatory compliance, among others (Wu et al., 2023). Alternatively, open-source models Meta9s Llama and Mistral9s models are available for any user to run and fine-tune on their computers (Touvron et al., 2023a); (Jiang et al., 2023). While some are categorized as "open-weight" due to some restrictions on their training details, they offer far more flexibility and accessibility than closed-source alternatives.

2.2 The Evolution and Architecture of Large Language Models

Large Language Models are a result of decades of research in neural networks, machine learning, and natural language processing within the bigger field of deep learning and artificial intelligence. (Zhang & Zong, 2019). The chronological progression illustrated in Figure 1. aims to point significant historical moments that influenced state-of-the-art LLMs we know today. The earliest AI program started in 1952 with the creation of Strachey's checkers, a program able to play a game of checkers, setting the stage for future developments in AI-based logic systems. (Strachey 1966).

But it was only in 1956, during the Dartmouth Conference the the term <artificial intelligence= was conceived, providing the starting groundwork for the formalization of the field (Staunton et al., 2019). Not long after, Rosenblatt (1958) introduced the Perceptron, a simple binary classification model that laid the foundation for neural networks by demonstrating how machines could be trained to classify patterns (Rosenblatt, 1958).The second big advancements in this same field came in 1986 with the development of Backpropagation by Rumelhart, Hinton, and Williams, which allowed neural networks to adjust their weights based on errors in their predictions (Rumelhart, Hinton, & Williams, 1986). This made it possible to train deeper, more complex networks making it possible for applications in healthcare during the 1990s where they were used to diagnose coronary artery disease (Baxt, 1990). As computational capacities soared and large datasets became more accessible (Anexo 1), neural networks began to outperform traditional machine learning models in tasks such as image and speech recognition. However, processing sequential data remained a barrier. Until the the introduction of Long Short-Term Memory (LSTM) networks by Hochreiter and Schmidhuber in 1997, overcame the problem of NN <remembering= long sequences of data (Hochreiter & Schmidhuber, 1997). These breakthroughs reflect the processing time-series data and text of modern language models. The development Word2Vec in 2013, also enabled the efficient embedding of words into numbers, helping computers recognize similarities and differences in meaning more efficiently (Mikolov et al., 2013).

Yet the true breakthrough in language modeling came in 2017 with the introduction of the Transformer model by Vaswani et al. (2017). Unlike previous NNs and LSTMs, that process information step by step, the Transformer uses what is called a "self-attention mechanism." allowing it to process all elements of a sequence simultaneously rather than sequentially. This innovation skyrocketed the efficiency and scalability of language models, enabling them to capture long-range dependencies in text more effectively, becoming the backbone of state-of-the-art LLMs like GPT, BERT, capable of performing tasks such as question answering and document classification (Vaswani et al., 2017). Following this innovation, OpenAI's GPT-3, launched in 2020, becoming one of the most influential LLMs, pushing the boundaries of natural language generation. GPT-3 was capable of producing highly coherent and contextually relevant text by leveraging its 175 billion parameters, demonstrating the effectiveness of large-scale pretraining and fine-tuning on diverse text corpora (Brown et al., 2020), and adapt to numerous applications beyond NLP, including medicine (Chen et al., 2023) and finance (Yang et al., 2023). The commercialization of LLMs followed shortly after, with the rise of open-

source initiatives and commercial applications in 2021. During this time, LLMs quickly spread to a wide range of industries, from mental health diagnostics to customer service automation, demonstrating the enormous potential of these models to solve practical issues. In the end, decades of progress in Deep Learning, Neural Networks, and Artificial Intelligence have led to the development of LLMs. The latest Transformer-based architectures and the early perceptron models are examples of how years of machine learning research and development have culminated in LLMs. In many fields, including mental health care, where their potential to analyze language and offer diagnostic support is still being fully realized, their unprecedented scale of text processing and generation positions them as transformative tools.

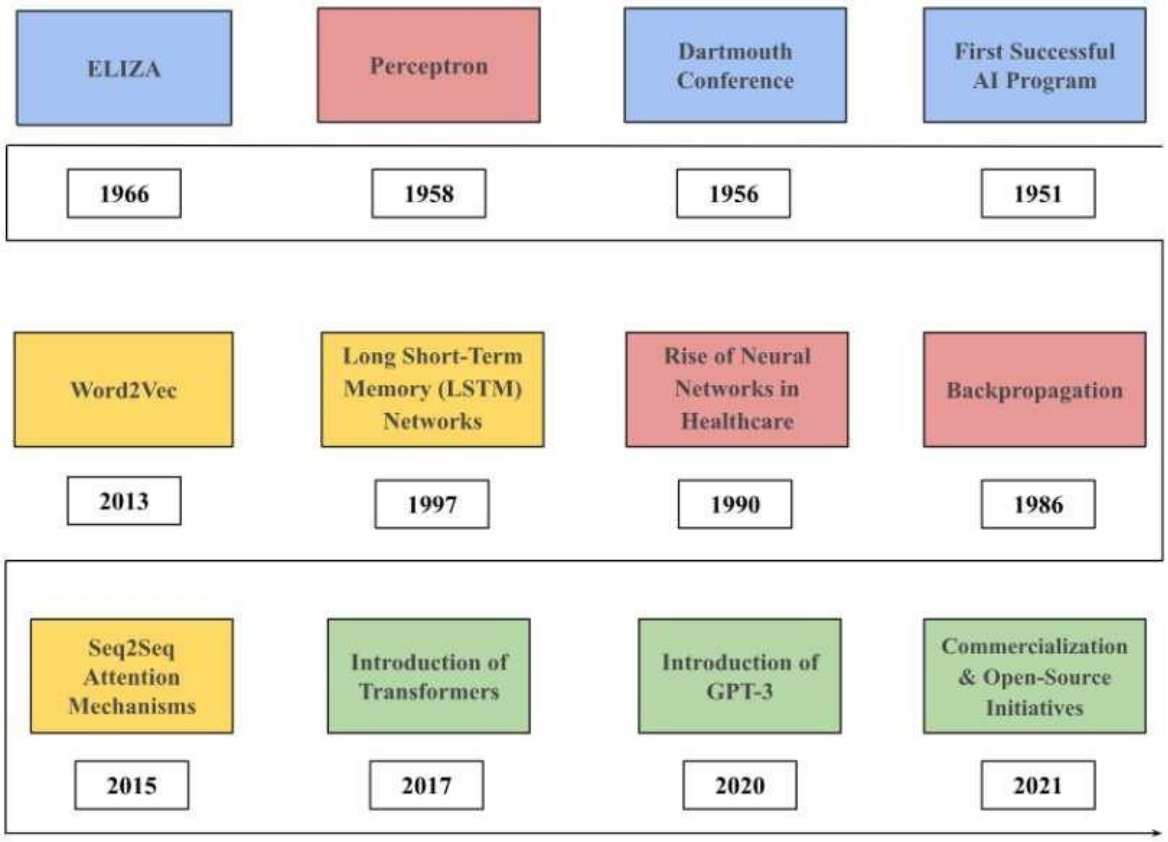


Figure 1: *Chronological progression of foundational events leading to LLMs, distinguished by colors: Neural Networks (NN) in red, Deep Learning (DL) in yellow, and Large Language Models (LLM) in green, Artificial Intelligence in blue.*

2.3 Mental Health: Challenges and the Role of LLMs

Mental health disorders impact approximately 970 million individuals globally and have costs exceeding $1 trillion each year (WHO, 2019), exceeding the economic impact of cancer, cardiovascular diseases, and diabetes. Given the importance of one9s psychological and emotional well-being, to an overall functional society, the United Nations Sustainable Development Goals (SDGs) (2022) recognized mental health as a significant public health priority.

With this recognition challenges related to early diagnoses must be assed. Traditional approaches primarily rely on self-reported information and standardized tools, such as the Depression Anxiety and Stress Scale and the Perceived Stress Scale. Although these methods are validated, they exhibit subjectivity, delays in diagnosis, and dependence on patient memory, rendering them inadequate for early detection. Mental health conditions frequently present in an episodic and comorbid manner, complicating diagnosis and treatment efforts. Diagnosing conditions such as bipolar disorder may require up to 10 years, frequently leading to insufficient treatment and worsened symptoms. (Dunstan et al., 2017); (Sunderland et al., 2019)

Large language models possess transformative potential to overcome the limitations of conventional mental health assessments through their ability to process and generate text at scale. These models can enhance mental health services via virtual platforms, promote early diagnosis, and allow for personalised treatment strategies by analysing extensive text data to identify patterns and insights essential for mental health professionals (Brown et al., 2020; Devlin et al., 2019). Examining the frequency, sentiment, and complexity of a patient's language can enable LLMs to identify early signs of conditions like depression or anxiety, thereby enhancing their utility in telemedicine. In this context, LLMs facilitate remote diagnostics by offering real-time insights that enhance traditional assessments. Additionally, LLMs improve therapeutic interventions by producing tailored responses in cognitive behavioural therapy (CBT) sessions, thereby reinforcing the therapeutic alliance and enhancing treatment outcomes (Shatte et al., 2019). This thesis examines two primary categories where large language models may significantly impact mental health: Clinical Applications and Data Analysis. The identified categories delineate the significant domains where LLMs can impact mental health care, either by directly aiding clinicians and patients or by analysing large datasets to yield meaningful insights.

2.3.1 Clinical Applications in Mental Health

Clinical applications in the context of mental health involves the administration of diagnostic tools, therapeutic approaches, and use of technologies to treat mental conditions such as depression, anxiety, and bipolar disorder. However, there are reoccurring challenges in current practices, for instance delayed diagnosis, where conditions like bipolar disorder can take up to a decade to diagnose accurately (Hirschfeld et al., 2003). Moreover, many diagnostic methods often rely on self-reported data, introducing bias and subjectivity to the matter, leading to potential misdiagnosis (American Psychiatric Association, 2013).

The lack of resources also strains mental health care systems, reducing the time available for personalized care (Rehm & Shield, 2019). Additionally, treatment adherence is a common challenge, as patients may not follow through with medication or therapy, leading to relapse and worsened symptoms (Emsley R. (2013). LLMs can have a positive influence to these issues by analyzing patient communication, therefore detect early signs of mental health conditions such as depression and anxiety, improving timely diagnosis (Torous et al., 2018). These models reduce subjectivity by objectively analyzing large datasets, leading to more accurate assessments (Devlin et al., 2018). Additionally, LLMs can be integrated into telemedicine platforms, providing scalable mental health support and easing the burden on clinicians (Munmun et al., 2023). In therapy, LLMs can offer personalized interventions, improving patient engagement and treatment adherence (Shatte et al., 2019). By addressing the core challenges of delayed diagnosis, diagnostic bias, resource constraints, and treatment adherence, LLMs hold the potential to significantly enhance clinical mental health care.

2.3.2 Data Analysis in Mental Health

The reviewed studies illustrate the increasing influence of Large Language Models (LLMs) in mental health, attributed to their capacity for analysing large datasets and providing significant insights. These models are revolutionising mental health care through early detection, risk identification, and data-driven decision-making, thereby offering essential support to clinicians and public health professionals.

The studies analysed consistently highlight the potential of LLMs to improve early detection of mental health disorders. Models such as Mental-Alpaca and FLAN-T5 demonstrate significant accuracy in detecting early indicators of conditions including depression, anxiety, and suicidal

ideation. Through the analysis of linguistic markers in patient communications, these models can identify subtle language shifts that may indicate the onset of mental health issues, often revealing patterns that traditional screening methods may overlook. The capacity to deliver early warnings renders LLMs especially beneficial in preventive care, as prompt intervention can greatly enhance patient outcomes.

One notable strength of LLMs identified in the studies is their capacity to integrate multimodal data, synthesising diverse sources including clinical records, social media activity, and real-time patient interactions. This approach allows LLMs to produce a more thorough understanding of an individual's mental health. Studies by Levkovich & Elyoseph (2023) and Mazumdar et al. (2024) illustrate that LLMs are capable of processing various behavioural, linguistic, and clinical data to provide personalised insights. This multimodal analysis enhances the accuracy and individualisation of mental health assessments, equipping clinicians with a comprehensive context for informed decision-making regarding treatment and care strategies. Beyond clinical applications, LLMs demonstrate potential in generating insights at the population level, especially in evaluating mental health trends among large cohorts. Models such as HeLM have been utilised to examine overarching trends in mental health data, providing significant insights into the manifestation of mental health conditions across various demographics and regions. Qi et al. (2023) investigated the capacity of LLMs to evaluate the mental health effects of global events, including the COVID-19 pandemic, through the analysis of changes in social media discourse. Population-level insights are essential for public health professionals, enabling the design of more effective interventions and the strategic allocation of resources.

Large language models have demonstrated efficacy in risk assessment, especially in pinpointing individuals at elevated risk for mental health crises. Numerous studies, such as those conducted by Qi et al. (2023), demonstrate that LLMs are capable of analysing language patterns to identify indicators of suicidal ideation or other crises. These models enhance protection by identifying high-risk individuals through subtle linguistic cues, facilitating earlier interventions and potentially saving lives. Nonetheless, the studies indicate that although LLMs are capable of identifying risks, human oversight is essential, particularly in high-stakes scenarios where the repercussions of errors can be significant.

2.4 Ethical and Practical Considerations

The integration of LLMs into mental health care holds ethical and practical challenges. First is the protection of data privacy and security due to the protective and sensitive nature of healthcare information. Regulations such as the Health Insurance Portability and Accountability Act (HIPAA) in the United States and the General Data Protection Regulation (GDPR) in the European Union enforce strict regulations and protocols to companies for the protection of patient data and the maintenance of confidentiality. HIPAA protocols include privacy and security measures that qualify patients to control their health information and establish limitations on the use and disclosure of such data. Lee and Lee (2008) illustrate that the security regulations of HIPAA is presented on cryptographic mechanisms in order to safeguard data integrity and confidentiality, thus facilitating the secure management of sensitive health information. Secondly, LLMs trained on large datasets have the possibility to acquire biases inherent to the same data, including those related to race, gender, or socioeconomic status, resulting in inequitable or imprecise evaluations in mental health care. If an LLM's training data primarily reflects a specific demographic, its performance in diagnosing or treating individuals from diverse backgrounds may be suboptimal, thereby worsening existing health disparities. To address these biases, it is essential to select training datasets meticulously, continuously evaluate model performance across various populations, and employ fairness-aware algorithms to reduce bias (Mehrabi et al., 2021).Furthermore, the management of sensitive health data is governed by stringent regulations, and large language models, which frequently handle extensive datasets, are required to adhere to legal frameworks that regulate data collection, processing, and storage. Failure to comply with HIPAA, GDPR, or comparable regulations may lead to legal consequences and could erode public confidence in AI-driven healthcare solutions (Dove, 2018). The increasing prevalence of LLMs in mental health care necessitates transparency in their application, including comprehensive documentation regarding the utilisation of patient data to ensure adherence to ethical and legal standards (Veale & Binns, 2017). A significant concern is the transparency and explainability of large language models (LLMs). The complexity of these models can render their decision-making processes challenging to interpret, potentially resulting in mistrust among clinicians and patients. Enhancing the interpretability of models and delivering clear explanations for AI-generated outputs are crucial for building trust and facilitating the safe and effective integration of LLMs in mental health care environments (Doshi-Velez & Kim, 2017).

3 Methodology

This study seeks to conduct a scooping review of the evidence regarding the application of Large Language Models (LLMs) in mental health, focussing on clinical applications and data analysis. The emphasis is on delineating application areas, categorising employed methods, and specifying sub-areas within mental health. This review aims to identify the challenges addressed by these models, including early detection of mental health disorders, therapeutic interventions, and personalised treatment planning, while assessing the value and utility of LLMs in these areas.

The goal is to answer several key research questions:

1. What are the primary areas of application for LLMs in mental health?
2. How are LLMs contributing to addressing significant challenges in mental health care?
3. What sub-areas of mental health care benefit the most from LLM integration?

These questions guided the search for relevant literature, helping to focus the review on the practical utility and results demonstrated by LLMs in the field.

3.1 Search Strategies

A scoping review of relevant studies was conducted to examine the current applications of LLM systems in mental health, focussing specifically on clinical applications and data analysis, beginning in May 2024. This approach sought to encompass the extensive research on LLMs in mental health, emphasising their application in tackling particular challenges within the discipline.

The investigation encompassed five databases: PubMed, IEEE Xplore, Scopus, Web of Science, and JMIR. The selected databases provide extensive coverage of medical, psychiatric, and AI-related literature, ensuring that the review addresses both technical advancements in LLM development and their practical applications in mental health.

3.1.1 Search Terms:

The search strategy integrated key terms pertinent to both LLMs and mental health to ensure comprehensive coverage. The terminology employed comprised:

Terms related to LLM include "ChatGPT." Large language models are a significant advancement in natural language processing within the field of artificial intelligence. These models utilise predictive algorithms to analyse and generate human-like text, enhancing various applications across multiple domains.
Mental health terminology includes: <Mental Health,= <Psychiatry,= <Mental Illness,= <Depression,= <Therapeutic Interventions,= <Bipolar Disorder,= and <Schizophrenia.=
The selected keywords aim to facilitate the identification of studies related to LLM applications throughout the entire range of mental health care, encompassing diagnostic tools and data-driven analysis.

3.2 Selection of Studies and Criteria for Eligibility

The author of this thesis conducted a selection of studies, emphasising the relevance of the literature regarding the application of LLMs in mental health and their role in addressing significant challenges within the field. The review sought to compile evidence regarding the utility of LLMs in clinical and data-driven contexts, focussing on their application in tasks including early diagnosis, treatment planning, and risk detection.

The inclusion and exclusion criteria applied aimed to ensure that the selected studies provided relevant empirical evidence.

Criteria for Inclusion

- Research focused on clinical applications, psychiatric care, and data analysis in mental health aims to provide a comprehensive understanding of the discipline.
- Research articles, preprints, and peer-reviewed papers addressing LLMs in mental health, clinical, or data analysis contexts, published in English from January 1, 2023, to August 2024, were included.

Studies conducted before 2023 were excluded due to the inadequacies of LLM models from that period, including GPT-2 and early iterations of GPT-3, which did not possess the requisite maturity, scalability, and contextual understanding for effective mental health applications. Previous models encountered ethical and practical constraints, such as bias, privacy issues, and restricted real-world efficacy in sensitive areas like mental health. Advancements in models such as GPT-4 in 2023 enhanced contextual understanding and ethical management of sensitive data, thereby increasing the viability of LLMs for clinical applications, including diagnosis and treatment planning. This review emphasises studies from 2023 onwards, aligning with the latest advancements in the field and ensuring relevance to current applications of LLMs in mental health care.

Criteria for Exclusion

- Editorials, opinion pieces, and studies lacking substantial evidence or empirical data were excluded.
- Articles that did not provide adequate detail regarding the methodology or model application were excluded from the review.

Due to the rapidly advancing field of LLM research, especially regarding mental health applications, preprints and non-peer-reviewed studies were taken into account. These studies frequently embody advanced developments that have not yet undergone formal publication processes but are essential for comprehending emerging trends in LLM applications. Incorporating these studies guarantees that the review reflects recent advancements and current discussions in the field, while also allowing for adaptability in identifying the most pertinent and timely evidence.

3.3 PRISMA

The study selection process followed a structured approach. The initial search across five databases resulted in approximately 1,000 articles. After removing duplicates, a total of 800 unique papers were evaluated based on their titles and abstracts. Of these, 75 full-text articles were assessed for eligibility, with 29 studies ultimately satisfying the inclusion criteria.

The studies were classified into two primary categories:

Clinical Applications: Research examining the application of LLMs in diagnosis, treatment planning, and therapeutic interventions within the field of mental health.

Research on the application of LLMs in analysing extensive datasets for early detection, risk prediction, and the generation of comprehensive insights into mental health trends.

The PRISMA diagram promotes transparency in the selection process and underscores the systematic methodology employed to include only pertinent, high-quality studies in the final review.

3.4 Data Extraction

Data extraction was conducted following a structured approach to ensure consistency and thoroughness. For each study, the key elements extracted included the author(s), year of publication, country or region, and the specific mental health case being examined. Special attention was given to the LLM models used, their applications within mental health settings (such as diagnosis, treatment recommendations, or data analysis), and the methodologies employed.

The sample sizes for each study were recorded to provide context regarding the scope of the research. Additionally, the main findings were extracted, focusing on the effectiveness, impact, and broader implications of LLMs in mental health applications. The main study designs4 whether experimental, observational, or comparative4were also documented to provide a comprehensive understanding of the methodologies used.

Limitations of each study were carefully noted, including potential weaknesses such as small sample sizes or methodological challenges. Ethical concerns, such as data privacy or model bias, were highlighted where applicable. Finally, the authors9 recommendations were captured to provide insight into suggested areas for future research, while predicted trends were noted to outline the anticipated future directions of LLM applications in mental health.

This structured data extraction approach allowed for the systematic gathering of relevant information, ensuring that the key aspects of each study were documented in alignment with the objectives of this thesis.

3.5 Qualitative Approach

A comprehensive analysis of the applications of LLMs in mental health was conducted using qualitative methods. This approach allowed for a more nuanced understanding of the efficacy of LLMs in this field by examining thematic patterns and contextual insights across the selected studies. The qualitative analysis explored the wider implications of LLM use in mental health, identifying recurring themes such as the models' capacity to replicate empathy, manage complex psychiatric discussions, and integrate into therapeutic environments. Additionally, expert evaluations and subjective assessments from the studies provided valuable insights into the practical applications and limitations of LLMs in clinical settings.
This narrative synthesis offered a deeper understanding of the opportunities and challenges involved in implementing LLMs in mental health care. The analysis examined the ethical considerations, practical applications, and broader impacts of LLMs on mental health services, offering a well-rounded perspective on their potential contributions and limitations.

4 Data

Table 1: *Summary of the 14 selected articles from the literature on* ***LLMs in mental health in Clinical Settings***

Ref.	Cases	Models	Methodology Used	Main Outcomes
Filienko et al., 2024	Comparison of prompting techniques to improve GPT-delivered Problem-Solving Therapy (PST).	GPT-3.5, GPT-4, InstructGPT (a fine-tuned version of GPT models)	The study compared prompting methods to improve LLM performance in delivering protocolized PST, specifically focusing on symptom identification and goal setting. Evaluations were based on automatic metrics and expert assessments.	Prompt engineering improved GPT models' consistency, quality, and empathy in delivering PST. However, limitations remain, and further refinement is needed for clinical reliability.
Lai et al., 2024	The Psy-LLM framework was developed to address the demand for mental health services using AI-based large language models in online psychological consultation settings.	Psy-LLM pre-trained on PanGu and WenZhong	The framework was evaluated using both intrinsic (perplexity) and extrinsic (helpfulness, fluency, relevance) metrics with human assessments.	Psy-LLM effectively generated coherent, relevant responses for mental health consultations, showing potential in alleviating professional shortages, though further improvements are needed.
Franco D'Souza et al., 2023	Evaluation of ChatGPT-3.5 in answering clinical vignettes related to psychiatry.	GPT-3.5	ChatGPT-3.5 was tested on 100 psychiatric vignettes. Expert psychiatrists evaluated its responses across 10 categories, with results summarized in graphs and tables.	ChatGPT-3.5 showed promise in answering psychiatric cases, but its reliability for clinical use remains insufficient.
Levkovich & Elyoseph, 2023	Comparison of ChatGPT-3.5 and ChatGPT-4 in evaluating depression and recommending treatment protocols against primary care physicians.	ChatGPT-3.5 and ChatGPT-4.	ChatGPT was assessed using vignettes depicting patients with varying depression severity and characteristics like gender and socioeconomic status. Responses were repeated 10 times to ensure consistency and evaluated against those of primary care physicians.	Both models demonstrated potential in adjusting treatment for mild and severe depression, though concerns remain regarding biases in therapeutic recommendations.

Sezgin et al., 2023	Assessment of clinical accuracy in responses to postpartum depression (PPD) questions from ChatGPT-4, LaMDA (Bard), and Google Search.	ChatGPT-4, LaMDA (Bard)	The models were queried with 14 PPD-related questions. Responses were evaluated by physicians against American College of Obstetricians and Gynecologists (ACOG) guidelines using a quality rating scale.	ChatGPT-4 provided the most clinically accurate responses, outperforming both Bard and Google Search in quality.
Hua et al., 2024	A scoping review of large language models (LLMs) used in mental health care, with human-assessed generative tasks.	GPT-3, GPT-3.5, GPT-4.	A review of 726 studies from databases like APA PsycNet, Scopus, and PubMed. Only 17 studies with human participants in real-world mental health scenarios were included, covering clinical assistance, counseling, therapy, and emotional support.	LLMs show potential in mental health care but lack standardized evaluation methods, raising concerns about privacy, transparency, and safety.
Kim et al., 2023	Development and evaluation of MindfulDiary, a journaling app powered by GPT-4 to support psychiatric patients.	GPT-4	A four-week field study was conducted with 28 patients with major depressive disorder and 5 psychiatrists. The app facilitated daily journaling through natural conversations, which were summarized and provided to psychiatrists via a dashboard.	MindfulDiary helped patients enhance daily journaling, and psychiatrists gained better insights into patient conditions, improving empathy and treatment planning.
Wang et al., 2023	Development of ClinicalGPT, a large language model fine-tuned for medical applications.	ClinicalGPT (fine-tuned from BLOOM-7B).	ClinicalGPT was fine-tuned using diverse medical datasets, including medical conversations, question-answering, and electronic health records (EHR). The model was evaluated using metrics like BLEU and ROUGE for medical conversation and multiple-choice accuracy for diagnostic tasks.	ClinicalGPT outperformed other models like ChatGLM-6B and LLAMA-7B in medical dialogue quality and medical exam accuracy, showing promise in clinical diagnosis and medical conversations.

Galatzer-Levy et al., 2023	Evaluation of Med-PaLM 2, a medical LLM, for predicting psychiatric functioning from patient interviews and clinical descriptions.	Med-PaLM 2, trained on a large corpus of medical knowledge	The model was tested on depression, PTSD assessments, and clinical case studies across various psychiatric disorders. Prompts were used to estimate clinical scores and diagnoses.	Med-PaLM 2 achieved high accuracy (0.80-0.84) in predicting depression scores, with results comparable to human clinical raters. The model showed promise in flexibly predicting psychiatric risk based on free-form descriptions.
Chen et al., 2023	Introduction of Diagnosis of Thought (DoT) prompting to detect cognitive distortions in psychotherapy using LLMs.	ChatGPT (GPT-3.5-turbo), GPT-4, Vicuna.	LLMs were guided through three stages—subjectivity assessment, contrastive reasoning, and schema analysis—to identify cognitive distortions. The models were evaluated on distortion detection and classification, with human expert validation.	DoT prompting significantly improved ChatGPT's performance on distortion detection, with GPT-4 surpassing full-training models in classification accuracy. Human experts validated the quality of diagnosis rationales.
Cho et al., 2023	Evaluation of LLMs for interactive language therapy for high-functioning autistic adolescents.	Custom LLM based on GPT-3.5.	Clinical psychologists and psychiatrists evaluated LLM interactions with hypothetical scenarios, focusing on empathy, communication, adaptability, and therapeutic alliance, using a custom scorecard.	LLM demonstrated strong empathy and adaptability, though it struggled with personalization and deep therapeutic connections. It shows potential as a supplementary tool for therapy but needs further refinement.
Ma et al., 2023	Study of the benefits and challenges of using LLM-based conversational agents like Replika for mental well-being support.	Replika (AI-based CA powered by GPT-3)	Qualitative analysis of 120 posts and 2,917 user comments from a mental health support subreddit. The study explored user experiences with the Replika app.	Replika provided non-judgmental, on-demand support, boosting self-confidence, but faced issues with harmful content, communication consistency, and overdependence, emphasizing the need for responsible implementation.

Loh & Raamkumar, 2023	Evaluation of large language models (LLMs) for generating empathetic responses in online mental health counseling.	GPT-3.5, GPT-4, Vicuna FastChat-T5, PaLM-2, Falcon-7B-Instruct.	LLMs were tested on the Empathetic Dialogues dataset, comparing their empathetic response generation with traditional dialogue systems. Models were evaluated on three empathy-related metrics: emotional reaction, interpretation, and exploration.	LLMs, particularly GPT-4 and VicunaT5, demonstrated superior empathetic response generation compared to traditional systems, showing potential in mental health counseling scenarios.
Heston, 2023	Evaluation of ChatGPT-3.5's responses to escalating prompts assessing mental health risks, including depression and suicidality.	ChatGPT-3.5	Twenty-five publicly available ChatGPT-3.5 agents were evaluated using two sets of escalating prompts (depression severity and PHQ-9 tool). The study measured when the agents referred users to human intervention and at what point the conversation stopped for recommending human intervention.	ChatGPT agents referred users at moderate depression levels but delayed firm intervention recommendations until severe depression prompts. Only two agents provided suicide hotline information, highlighting the need for stronger safety mechanisms before clinical use.

Table 2: *Summary of the 15 selected articles from the literature on LLMs in mental health in* ***Data Analysis***

Author(s)	Case	Model	Methodology Used	Main Outcomes
Adhikary et al., 2024	Evaluation of large language models (LLMs) in summarizing therapy sessions for mental health counseling.	BART, T5 (Text-To-Text Transfer Transformer), GPT-2, GPT-Neo, GPT-J, FLAN-T5, Mistral, MentalBART, MentalLlama, Llama-2, Phi-2	This study created a dataset of 191 counseling sessions with summaries focused on 3 counseling components. The models were evaluated on counseling-component–guided summarization using ROUGE and BERTScore metrics. Mental health professionals conducted a qualitative evaluation across 6 parameters.	Task-specific models, especially Mistral, outperformed others in expert evaluations and automatic metrics. However, improvement is needed in opportunity costs and perceived effectiveness before clinical application.
Xu et al., 2024	Evaluation of multiple LLMs for mental health prediction tasks using online text data.	Alpaca, Alpaca-LoRA, FLAN-T5, GPT-3.5, GPT-4, Mental-Alpaca, and Mental-FLAN-T5.	The study performed experiments with zero-shot, few-shot, and instruction-finetuning setups on social media datasets. Models were evaluated for their accuracy in mental health predictions.	Instruction-finetuned models (Mental-Alpaca, Mental-FLAN-T5) outperformed GPT-3.5 and GPT-4 in accuracy by 10.9% and 4.8%, respectively, showing promising potential for mental health tasks. However, ethical concerns and biases remain
Mazumdar et al., 2024	Introduction of GPTFX, an AI-based framework for mental health disorder detection and explanation generation.	GPT-3 with fine-tuning for classification and explanation generation.	GPT embeddings combined with machine learning models for classifying mental health disorders, and fine-tuning of GPT-3 for generating explanations. The framework was tested for real-time monitoring in AI-IoMT devices.	GPTFX achieved 87% accuracy in classification and a Rouge-L score of 0.75, outperforming traditional algorithms in reliability for real-time mental health monitoring.
Levkovich & Elyoseph, 2024	Evaluation of ChatGPT-3.5 and ChatGPT-4 in assessing suicide risk through vignette studies.	ChatGPT-3.5, ChatGPT-4	ChatGPT models were evaluated using suicide risk factors (perceived burdensomeness and thwarted belongingness) from clinical vignettes, with results compared to mental health professionals' assessments.	ChatGPT-4 closely matched professional assessments, while ChatGPT-3.5 underestimated risks. ChatGPT-4 also overestimated psychache, highlighting the need for further refinement.

Lamichhane, 2023	Evaluation of GPT-3.5-turbo for mental health classification tasks.	ChatGPT-3.5-turbo.	ChatGPT was tested on stress, depression, and suicidality detection tasks using publicly available social media datasets. F1 scores were computed for each task to evaluate performance.	ChatGPT achieved F1 scores of 0.73 for stress detection, 0.86 for depression detection, and 0.37 for suicidality detection. It outperformed baseline models but underperformed on suicidality classification. Further fine-tuning could improve performance.
Liu et al., 2023	Evaluation of large language models (LLMs) as few-shot learners for health tasks using physiological and behavioral data.	PaLM (24 billion parameter transformer architecture).	LLMs were evaluated on tasks such as activity recognition, heart rate classification, calorie estimation, and mental health assessments using data from wearable devices. Models were tested with zero-shot, prompt tuning, and supervised baselines.	Prompt-tuned LLMs significantly outperformed zero-shot models and supervised baselines, showing promise as few-shot learners for health-related tasks like cardiac analysis and mental health screening.
Qi et al., 2023	Evaluation of large language models (LLMs) for suicide risk and cognitive distortion detection on Chinese social media.	GPT-3.5, GPT-4, ChatGLM2-6B, and GLM-130B.	Zero-shot, few-shot, and fine-tuning strategies were tested on annotated datasets from Chinese social media. The models were evaluated on their ability to detect suicide risk and cognitive distortions.	GPT-4 consistently outperformed GPT-3.5, especially after fine-tuning, in detecting cognitive distortions and suicide risk.
Englhardt et al., 2023	Evaluation of LLMs for analyzing mobile and wearable sensor data for mental health insights.	GPT-3.5, GPT-4, PaLM-2.	LLMs were tested on multi-sensor data (sleep, activity, phone usage) for binary depression classification using chain-of-thought prompting and reasoning with domain-specific data.	PaLM-2 achieved 61.1% accuracy in depression classification, outperforming GPT-4 and GPT-3.5. LLMs showed promise in generating clinical insights but require careful prompt design to avoid bias.

Jin et al., 2024	Evaluation of LLMs for mental health using the PsyEval benchmark, covering knowledge, diagnostic, and therapeutic tasks.	GPT-4, GPT-3.5-turbo, LLaMa, LLaMa2, Alpaca, Vicuna, ChatGLM2, and MOSS.	Eight models were tested across six tasks, including mental health question-answering, diagnosis from text and dialogue, therapeutic conversations, empathy, and safety.	GPT-4 performed best in mental health tasks, with challenges in empathy and complex diagnosis. All models require improvement in handling nuanced mental health scenarios.
Kerz et al., 2023	Development of explainable AI (XAI) models for mental health detection based on language behavior from social media.	BiLSTM and pretrained language models (PLMs) for mental health detection.	The study focused on interpreting mental health predictions using syntactic, lexical, and emotional features extracted from social media posts. XAI techniques like LIME and AGRAD were used to provide explanations for model predictions across five mental health conditions.	XAI approaches offered improved transparency in detecting mental health conditions, balancing interpretability and accuracy.
Jin et al., 2023	Introduction of TrialGPT, an LLM framework for matching patients to clinical trials.	TrialGPT	TrialGPT was evaluated on 184 patients and 18,000 trial annotations, predicting patient eligibility on a criterion-by-criterion basis. A user study assessed its impact on screening time in clinical trial matching.	TrialGPT achieved 87.3% accuracy on criterion-level predictions and reduced screening time by 42.6%, outperforming competing models by up to 57.2%.
Li et al., 2023	Investigation of LLMs for enhancing health data interoperability using FHIR (Fast Healthcare Interoperability Resources).	GPT-4	The model was tasked with converting clinical text snippets into FHIR resources, evaluated on 3,671 clinical text snippets. Performance was measured based on accuracy in converting text to FHIR standards.	The LLM achieved over 90% accuracy in exact matches compared to human annotations, demonstrating potential for streamlining clinical data interoperability in healthcare.

Binz & Schulz, 2023	Transforming large language models (LLMs) into cognitive models by fine-tuning them with data from psychological experiments.	LLaMA (65B parameters).	LLaMA was fine-tuned on behavioral datasets, including decision-making tasks (decisions from descriptions and experience). Human choices were predicted based on the model's embeddings, and individual differences were captured using random effects.	Fine-tuned models outperformed traditional cognitive models, accurately simulating human-like behavior in decision-making tasks and generalizing to unseen tasks.
Yang et al., 2023	Exploring large language models for interpretable mental health analysis, particularly in detecting and explaining mental health conditions.	GPT-3.5-turbo, InstructGPT-3, LLaMA-7B, LLaMA-13B.	The study tested models across 11 datasets for tasks such as mental health condition detection, cause detection, and emotion recognition, employing zero-shot and emotion-enhanced prompting. Human evaluations were used to assess explanation quality.	ChatGPT performed best but still lags behind task-specific methods. Prompt engineering with emotional cues improved performance, and ChatGPT generated near-human-level explanations.
Belyaeva et al., 2023	Development of HeLM, a multimodal large language model for health grounded in individual-specific data.	HeLM (Health Large Language Model) based on Flan-PaLMChilla 62b.	HeLM was evaluated on data from the UK Biobank, combining clinical, demographic, and high-dimensional time-series data (e.g., spirometry) for disease risk estimation. The model used encoders to map non-text data into token embedding space.	HeLM outperformed traditional models like logistic regression and XGBoost in predicting disease risk for conditions such as asthma, showing promise in personalized healthcare applications.

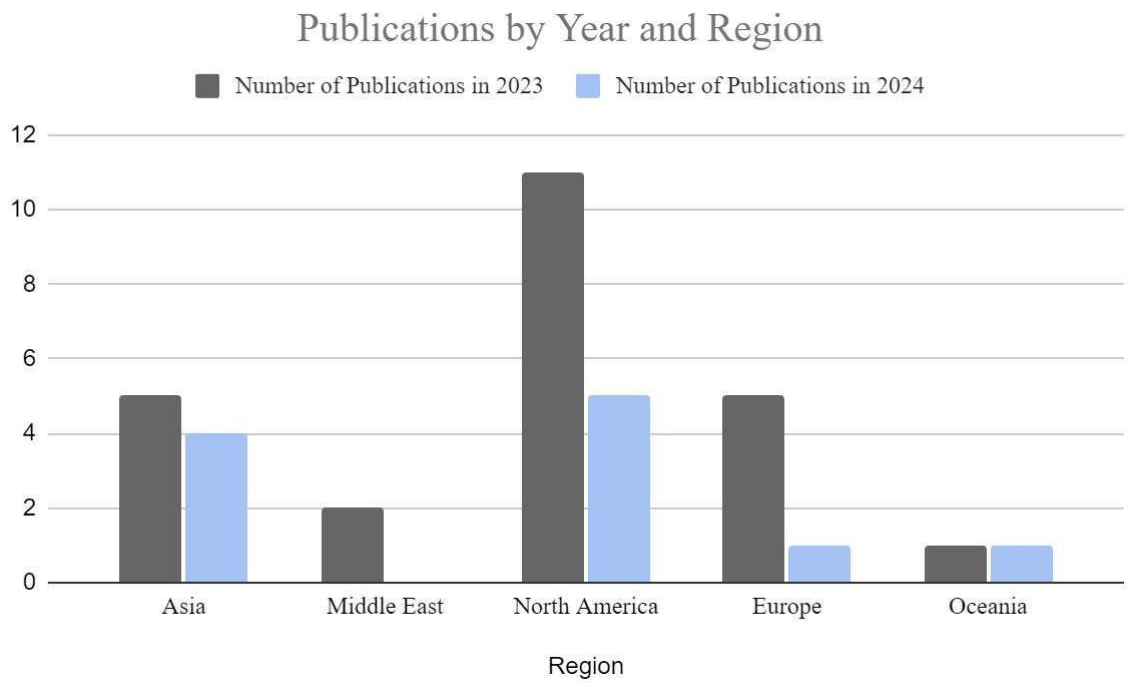


Figure 2: Publications by Year and Region

Exhibit 1: Proposed Taxonomy of LLM Applications in Mental Health

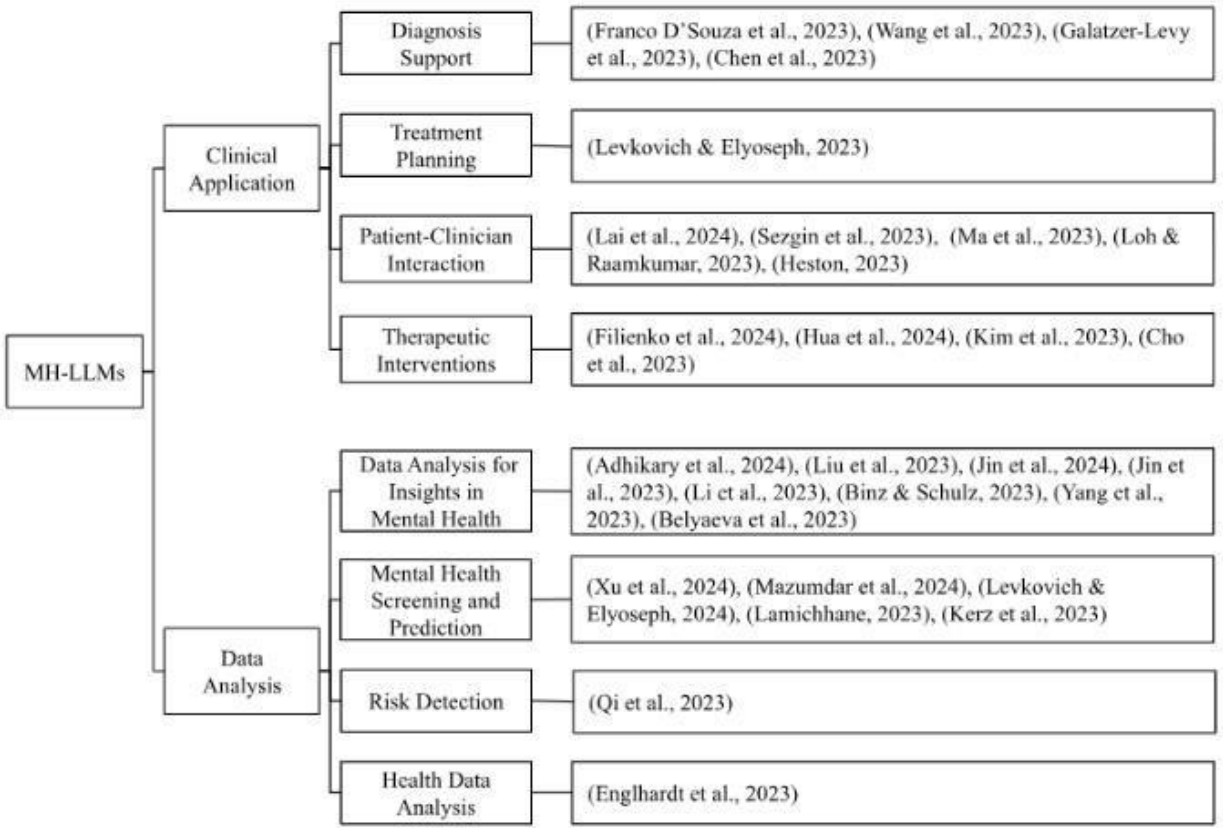


Figure 3: Taxonomy of LLM Applications in Mental Health. Legend:

***Diagnosis Support**: The use of LLMs to assist in diagnosing mental health conditions by analyzing patient data and identifying relevant patterns.*

***Treatment Planning**: LLMs provide recommendations for personalized treatment strategies based on patient history and data analysis.*

***Patient-Clinician Interaction**: Enhancing communication between patients and clinicians, LLMs offer real-time session summaries and aid in tracking patient progress.*

***Therapeutic Interventions**: Conversational agents powered by LLMs simulate therapy, offering support and therapeutic advice in mental health contexts.*

***Data Analysis in Mental Health**: LLMs analyze extensive mental health datasets to identify trends and generate insights that aid in research and patient care.*

***Mental Health Screening and Prediction**: LLMs screen for potential mental health risks and predict the onset of disorders based on behavioral and linguistic data.*

***Risk Detection**: Advanced algorithms detect high-risk behavior, such as signs of suicidality or severe mental health issues.*

***Health Data Analysis**: LLMs process health data to track population health trends and assess the effectiveness of treatments.*

Table 3: *SWOT Analysis of LLMs for Clinical Applications in Mental Health*

STRENGTHS (Effectiveness)	WEAKNESSES (Limitations)
Improved Patient-Clinician Interaction: Models like Psy-LLM have demonstrated effectiveness in generating coherent and relevant responses, helping bridge communication gaps in mental health settings (Lai et al., 2023) **Diagnostic Support and Treatment Recommendations:** ChatGPT-4 and similar models showed promising results in psychiatric diagnosis support, answering clinical questions with a high level of accuracy (Franco D'Souza et al., 2023; Sezgin et al., 2023). **Enhanced Treatment Planning**: LLMs such as ChatGPT-4 have shown effectiveness in providing comprehensive treatment recommendations for conditions like schizophrenia (Filienko et al., 2024).	**Inconsistent Clinical Accuracy:** Earlier models like GPT-3.5 exhibit inconsistencies in risk detection and treatment planning, sometimes offering inappropriate advice or overlooking critical cues (Levkovich & Elyoseph, 2023). **Limited Human Interaction Context:** LLMs struggle with capturing the full complexity of human emotions and situations, leading to potential gaps in clinical empathy (Lai et al., 2023). **Bias and Ethical Concerns:** Models may reflect biases inherent in their training data, leading to skewed clinical recommendations, especially for underrepresented patient populations (Sezgin et al., 2023).
OPPORTUNITIES (Trends)	**THREATS (Ethical Issues)**
Refining AI Models for Enhanced Risk Detection: There is growing interest in developing more sophisticated algorithms to improve LLMs' ability to detect mental health risks, including suicidality (Levkovich & Elyoseph, 2023). **Human-AI Collaboration**: LLMs offer a valuable tool for augmenting clinician expertise, particularly in diagnosis and treatment recommendations, allowing for more personalized and data-driven mental health care (Franco D'Souza et al., 2023). **Expanding Global Access to Mental Health Services:** As LLMs improve in multilingual capabilities, they can help expand access to mental health resources in underserved regions, contributing to global mental health equity (Lai et al., 2023).	**Over-reliance on LLMs**: There is a risk that clinicians may become overly reliant on AI-driven models, reducing their critical judgment in complex cases (Sezgin et al., 2023). **Ethical Challenges in Clinical Settings**: As LLMs become more integrated into clinical decision-making, there is an increasing concern about their ethical deployment, especially in high-stakes situations (Levkovich & Elyoseph, 2023). **Inappropriate or Harmful Advice**: If not closely monitored, LLMs can generate incorrect or harmful recommendations, particularly in mental health crises, raising concerns about their unsupervised use (Franco D'Souza et al., 2023)

Table 4: *SWOT Analysis of LLMs for Data Analysis in Mental Health:*

STRENGTHS (Effectiveness)	WEAKNESSES (Limitations)
High Accuracy in Data Analysis: Studies have shown that models like GPTFX and fine-tuned versions of GPT-3.5 achieved impressive accuracy rates (up to 87%) in classification and data analysis tasks, particularly in mental health contexts (Mazumdar et al., 2024). **Effective Screening and Prediction**: Fine-tuned LLMs, such as Mental-Alpaca and Alpaca-LoRA, have demonstrated significant potential in mental health screening, providing robust predictive capabilities (Xu et al., 2024). **Task-Specific Performance**: Task-focused LLMs, like Mistral, performed exceptionally well in handling specific mental health data analysis tasks, outperforming general-purpose models (Adhikary et al., 2024).	**Dependency on Quality of Training Data**: The performance of models is highly dependent on the quality and diversity of the datasets they are trained on, limiting their effectiveness in real-world scenarios if these datasets are not representative (Mazumdar et al., 2024). **Limited Contextual Understanding**: While LLMs excel in structured data analysis, they struggle with capturing the nuances of complex mental health cases where emotional context plays a significant role (Levkovich & Elyoseph, 2024). **Challenges in Real-World Deployment**: Despite their potential, models like Alpaca and FLAN-T5 face limitations when it comes to real-world application and deployment, particularly due to scalability and integration issues (Xu et al., 2024).
OPPORTUNITIES (Trends)	**THREATS (Ethical Issues)**
Integration with Clinical Tools: There is a growing opportunity to integrate LLMs with existing mental health diagnostic tools, augmenting clinical decision-making with data-driven insights (Adhikary et al., 2024). **Interpretable AI for Mental Health**: As the demand for interpretable AI systems increases, LLMs present a unique opportunity to offer transparent, explainable insights that can improve mental health assessments (Mazumdar et al., 2024). **Refinement of AI Models**: Further refinement of AI-driven data analysis frameworks presents an opportunity to enhance their accuracy, particularly by incorporating more diverse datasets and addressing biases (Xu et al., 2024).	**Bias in Data Analysis:** The inherent bias present in training data poses a significant threat to the fairness and accuracy of mental health insights generated by LLMs, potentially leading to skewed predictions and assessments (Levkovich & Elyoseph, 2024). **Ethical and Privacy Concerns:** The use of LLMs for mental health data analysis raises ethical concerns regarding patient data privacy and the transparency of AI-driven insights (Mazumdar et al., 2024). **Over-reliance on AI for Screening:** The increasing use of AI for mental health screening risks reducing human oversight in sensitive assessments, potentially leading to over-reliance on automated systems and neglect of nuanced clinical judgment (Levkovich & Elyoseph, 2024)

5 Discussion

5.1 Clinical Applications

The studies reviewed underscore the expanding role of LLMs in clinical mental health settings, particularly in supporting diagnostic processes, treatment planning, patient-clinician communication, and therapeutic interventions. A notable strength, as highlighted in multiple studies, is the high diagnostic accuracy of models such as GPT-4 and Psy-LLM in identifying complex mental health disorders, more specifically the ability to process patient data4ranging from conversations to medical histories4with a level of accuracy comparable to that of human clinicians. For instance, studies by Franco D9Souza et al. (2023) and Sezgin et al. (2023) show how LLMs can effectively perform diagnostic tasks in resource-constrained environments where mental health professionals are not readily available, providing a scalable solution to address mental health challenges in underserved areas.

Moreover, the scalability of these models presents a significant advantage. LLMs can automate the labor-intensive process of extracting insights from patient interactions, enabling more efficient screenings and facilitating broader access to mental health care globally. However, despite their strengths in diagnostic support, LLMs also face challenges. The SWOT analysis identifies inconsistent risk detection and inadequate crisis response as notable weaknesses. Levkovich & Elyoseph (2023) indicate that although LLMs perform well in controlled environments, they often struggle in real-time mental health crises. For example, ChatGPT-3.5 has difficulty recognizing and appropriately responding to high-risk situations, such as suicidal ideation, where immediate and nuanced intervention is crucial.

In the context of treatment planning, Levkovich & Elyoseph (2023) also highlight the potential of LLMs to generate evidence-based treatment recommendations by analyzing vast amounts of clinical data. These models provide valuable support for clinicians in developing personalized treatment plans, particularly by identifying patterns in patient histories and treatment outcomes. However, as the SWOT analysis notes, LLMs still fall short in capturing the emotional and contextual nuances essential to mental health care. Treatment plans generated by these models often lack the depth of understanding that human clinicians can provide, especially in emotionally complex or highly individualized cases. Therefore, while LLMs serve as useful tools in streamlining treatment planning, they should not be relied upon as standalone decision-makers.

LLMs have also been instrumental in enhancing patient-clinician interaction by summarizing patient sessions in real-time. This reduces the administrative burden on clinicians and allows them to focus more on patient care. For example, studies by Lai et al. (2024) and Sezgin et al. (2023) demonstrate how LLMs can offer immediate feedback during therapy sessions, thereby increasing the efficiency of care delivery. However, a critical limitation remains: LLMs struggle to interpret the emotional depth of conversations, a vital aspect of mental health care. This limitation raises concerns about the over-reliance on AI, as clinicians may defer too much to AI-generated summaries, potentially missing critical emotional cues better understood by human practitioners.

In therapeutic interventions, LLMs have shown potential in providing basic mental health support, particularly through conversational agents. Studies by Filienko et al. (2024) and Hua et al. (2024) explore how LLMs simulate empathetic dialogue, offering initial mental health assistance, especially in underserved regions. However, the SWOT analysis emphasizes ethical concerns, particularly regarding simulated empathy. Although LLMs can mimic empathetic dialogue, this simulation often feels artificial and lacks the depth needed for more complex therapeutic settings.

5.2 Data Analysis

LLMs are also play an important role in data analysis in mental health by leveraging their ability to process large datasets, which is transforming mental health screening, risk detection, and broader data-driven analysis. As demonstrated in Table 6.2, models like Mental-Alpaca and FLAN-T5 have achieved high accuracy in screening and prediction tasks, outperforming traditional models in both speed and precision (Xu et al., 2024). These models are particularly effective at identifying early signs of mental health conditions, such as depression and suicidal ideation, often identifying subtle patterns that might be overlooked by traditional methods.

The SWOT analysis highlights a key strength of LLMs: their ability to integrate multimodal data from various sources, such as behavioral, clinical, and even social media data, to offer more personalized mental health insights. This capability opens the door to more comprehensive mental health monitoring, allowing for real-time assessments of patients' conditions based on a wide range of inputs. However, the generalizability of LLMs across diverse populations remains a challenge. Studies by Levkovich & Elyoseph (2023) and Mazumdar et al. (2024) indicate that most models have been trained and tested on datasets from North America and Asia, raising concerns about their effectiveness in other regions with distinct

socio-cultural contexts. Ensuring the generalizability of these models to a wider range of populations will be critical for future research.

Another promising application of LLMs is in risk detection, particularly in identifying individuals at high risk for crises like suicidal ideation. For instance, Qi et al. (2023) examines how LLMs can analyze communication to detect early signs of crises, such as suicidal ideation, thereby enabling timely interventions. However, the SWOT analysis also underscores potential threats posed by the use of LLMs for risk detection, particularly regarding data privacy and ethical concerns. The integration of these technologies into mental health care must be followed by rigorous data protection guidelines to safeguard patient privacy and autonomy considering the nature of the data being stored. Mazumdar et al. (2024) have pointed out concerns about the ethical implications and potential risk of using LLMs for data-driven mental health insights, emphasizing the need for transparency and accountability in their deployment.

A proposed taxonomy for LLM applications in mental health is presented, considering the findings from the reviewed studies. This taxonomy categorizes LLM applications into two primary domains: Clinical Applications and Data Analysis. Within the domain of Clinical Applications, there are four subcategories where LLMS were utilized—Diagnosis Support, Treatment Planning, Patient-Clinician Interaction, and Therapeutic Interventions.

In the domain of Data Analysis, the studies analyzed were equally divided into four subcategories— Data Analysis in Mental Health, Mental Health Screening and Prediction, Risk Detection, and Health Data Analysis. This structured framework offers a clear overview of how LLMs are currently being used in mental health care, providing a foundation for future developments and research in this field. The taxonomy serves as a roadmap for understanding the diverse roles LLMs can play in enhancing mental health services, from supporting clinicians in diagnostic tasks to improving data-driven insights that can inform mental health policies and interventions. However, it also emphasizes the need for ongoing research to address the challenges and ethical concerns associated with LLM deployment in clinical settings, particularly in relation to crisis management and data privacy. By refining and expanding upon this taxonomy, future studies can contribute to the broader integration of LLMs in mental health care, ensuring that these tools are both effective and ethically sound.

6 Limitations

6.1 Geographic Distribution and Study Selection

A key limitation of this analysis is the geographic distribution of studies (Figure 6.3). The studies included in this analysis were curated based on the author9s selection criteria, focusing on those most relevant to the objectives of this thesis. While the selected studies provide a useful overview of research across regions such as North America, Asia, Europe, and the Middle East, it is important to note that this distribution does not necessarily reflect the global landscape of LLM research in mental health. The selection process inherently limits the analysis, as certain regions or research perspectives may have been unintentionally excluded. Therefore, any conclusions drawn about geographic trends should be interpreted cautiously, as the findings are not comprehensive or fully representative of global research efforts. The curation of studies also introduces a potential selection bias, which may influence the interpretation of regional contributions to LLM research. For example, the concentration of studies from North America, particularly the United States, may reflect both the prominence of AI research in that region and the focus of this thesis, rather than a true global leadership in LLM-based mental health applications. This limitation must be considered when discussing the geographic spread of LLM research and its broader implications.

6.2 Data Quality and Bias in Model Training

Hua et al., (2024) and Liu et al., (2023) point out the quality and bias of the data used to train LLMs as significant limitation. Many LLMs, such as GPT-4 and Psy-LLM, rely on vast datasets that may contain inherent biases, particularly related to race, gender, socioeconomic status, or cultural perspectives. This can lead to skewed mental health diagnoses or biased treatment recommendations, particularly for underrepresented groups.

Furthermore, the availability of high-quality datasets is constrained by data privacy regulations and the sensitive nature of mental health data. As discussed in the Data Analysis section, the success of LLMs in screening and predicting mental health conditions is closely tied to the comprehensiveness of the training data. Data sparsity or incomplete information can reduce the reliability of the models, and data privacy laws such as GDPR often limit the sharing of sensitive mental health data, restricting access to high-quality datasets that could improve model performance.

6.3 Methodological Constraints and Study Selection

The methodology used to assess the performance of LLMs in mental health contexts introduces its own set of limitations. While this thesis used state-of-the-art models and performance metrics (e.g., F1 scores, ROUGE scores), these quantitative measures often fail to capture the full complexity of mental health interactions. For example, while LLMs may score highly in diagnosing conditions based on textual input, these metrics do not account for the emotional depth and contextual understanding that are essential in mental health care. In the weaknesses section of the SWOT analysis, describes the difficulty of the models to interpret emotional nuances, as well as their lack of contextual awareness which could lead to misdiagnoses or inappropriate treatment recommendations. Furthermore, this limitation is particularly concerning in high-stakes situations, such as mental health crises, where a deeper understanding of the patient9s emotional state is critical. Additionally, the selection of studies, including the decision to incorporate preprints and non-peer-reviewed research, introduces further limitations. While this approach enabled the inclusion of the latest advancements and emerging trends in LLMs for mental health, it also comes with the risk that some studies may not have undergone the rigorous scrutiny of peer review. As preprints represent early-stage research, their findings are subject to change upon formal review and could potentially introduce speculative results or methodological flaws into the analysis. This flexibility was necessary to capture the real-time evolution of the field, but it also means that some findings drawn from these studies may lack stability or long-term credibility. In particular, the reliance on non-peer-reviewed studies limits the generalizability and reliability of some of the conclusions, as these studies may later be revised or retracted. This limitation showcases the trade-off between capturing the most current developments in the field and ensuring the accuracy of the research used in this thesis.

6.4 Ethical and Data Privacy Concerns

The use of LLMs in mental health raises significant challenged regarding ethical concerns, particularly around data privacy and bias in clinical recommendations. One of the primary threats posed by LLMs in the SWOT is their potential to generate biased or inappropriate advice, especially when dealing with sensitive mental health cases. These biases are often caused by the data used to train this models, which can disproportionately affect certain populations, potentially leading to misdiagnoses or harmful treatment recommendations.

The collection and process of mental health data, which is among the most sensitive types, increases the risk of data privacy violations, ensuring its protection is paramount. Although LLMs can offer significant benefits in improving mental health outcomes, the absence of clear ethical frameworks or regulatory oversight increases the risk of misuse or data breaches, further complicating the integration of these models into clinical practice.

6.5 Limitations in Crisis Management

LLMs' limitations are particularly pronounced in their ability to manage mental health crises. As presented in Levkovich & Elyoseph, (2024), models like ChatGPT-3.5 struggle with recognizing and responding to high-risk situations such as suicidal ideation or severe anxiety attacks. This inability to effectively manage crises is a critical limitation, especially as LLMs are increasingly deployed in therapeutic settings. While these models can offer general support in low-risk scenarios, their failure to handle acute crises highlights the need for human oversight in all clinical applications of LLMs. Without proper escalation mechanisms or intervention protocols, these models pose significant risks to patient safety.

6.6 Computational and Resource Constraints

Finally, the computational requirements for training and deploying LLMs present another limitation. The development of large-scale models like GPT-4 requires significant computational power, which limits the accessibility of these technologies in low-resource regions or smaller research institutions. The high cost and resource-intensive nature of training LLMs also pose barriers to their widespread adoption, exacerbating inequalities in access to advanced mental health tools. Moreover, the fine-tuning of models for specific mental health applications, while necessary for improving accuracy, further increases the computational burden, making it difficult for regions with fewer resources to benefit from these advancements.

While LLMs hold substantial promise for transforming mental health care, these limitations underscore the need for caution in their implementation. In order for these technologies to be deployed safely and equitably, concerns regarding data quality, bias, geographic representation, methodological constraints, and ethical concerns must be carefully addressed. The findings of this thesis provide a foundation for further research, but overcoming these limitations will be critical to the broader adoption and success of LLMs in global mental health systems.

7 Conclusions and Future Work

Conclusions

This thesis has explored the applications of Large Language Models (LLMs) in the field of mental health, specifically focusing on their roles in both clinical applications and data analysis. Through an in-depth analysis of 29 selected studies, the strengths, limitations, and ethical considerations surrounding the deployment of LLMs in mental health care were identified. The findings highlight that LLMs, such as GPT-4, Psy-LLM, and FLAN-T5, are emerging as powerful tools for diagnostic support, treatment planning, patient-clinician interaction, and mental health data analysis. These models have shown high diagnostic accuracy in identifying complex mental health conditions like schizophrenia and depression, positioning them as potential game-changers in areas with limited access to human clinicians. Moreover, the scalability of LLMs offers the possibility of reaching larger populations quickly and effectively, addressing a pressing global need for accessible mental health care.

On the other hand, this thesis has also underlined the limitations of LLMs, especially in handling high-stakes mental health crises and capturing the emotional and contextual depth of patient conversations. While LLMs can process vast amounts of textual data and provide recommendations based on it, their inability to fully comprehend the nuanced emotional states of patients could limit their effectiveness in certain therapeutic settings. Additionally, the decision to rely on preprints and non-peer-reviewed studies introduces risks in terms of data reliability and generalizability, hence reflecting the fast-developing but still unstable nature of LLM research in this domain.

The SWOT analysis aims to offer a comprehensive view of the current condition of LLMs in mental health, highlighting opportunities for innovation, such as improved crisis management algorithms and wider cultural adaptation, as well as the ethical threats, including biases in clinical recommendations and concerns regarding data privacy. These findings highlight the importance of using LLMs to enhance mental health care, but they should only be used as a support tool and not as a replacement for human involvement. It is crucial to maintain human oversight in any application of LLMs in clinical settings

Future Work

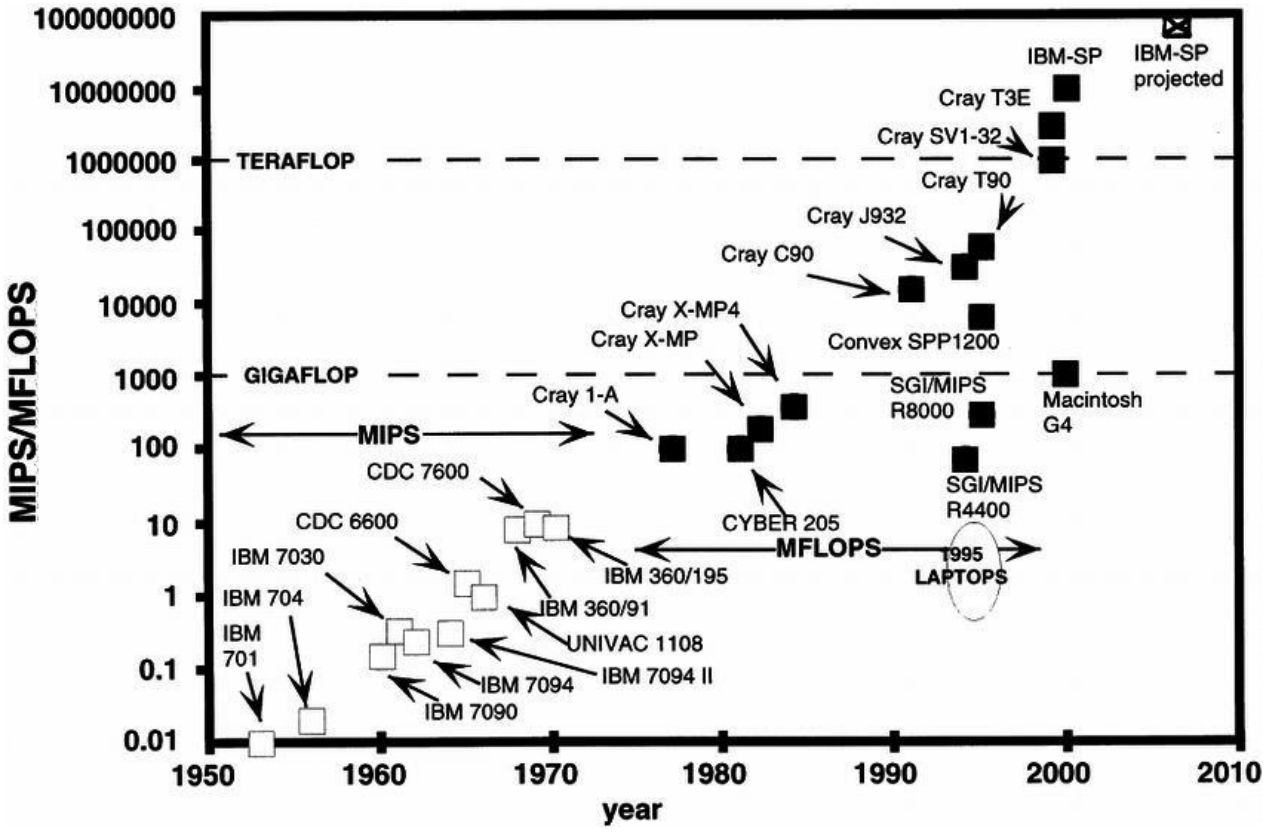
While this thesis has conducted a primary fundamental examination of the uses of LLMs in mental health, there are many key areas that need further investigation and development in order to enhance the efficacy and ethical deployment of these technologies. Firstly, while general-purpose models like GPT-3.5 and GPT-4 offer broad capabilities, models that are fine-tuned specifically for mental health applications such as Psy-LLM have shown potential to address the unique challenges of this field. Next focus on developing LLMs that are capable of integrating multimodal data[4] such as text, audio, and behavioral cues, in order to improve the contextual understanding of patient interactions. This could enable more personalized mental health care, and address the current limitations observed when dealing with complex therapeutic scenarios.

Moreover, further studies should aim to address and find solutions to the ethical challenges, particularly around bias and data privacy. As a way to reduce bias in LLM training, these models have to be trained in diverse and representative datasets, which include data from underrepresented populations, ensuring that results are applicable across different cultural and demographic contexts. Also, the continuous development and implementation of stronger data privacy frameworks, ensures that the sensitive nature of mental health data is safeguarded in all LLM applications.

Another important direction for future work is the improvement of crisis detection capabilities in LLMs. Current models have shown limitations in responding appropriately to mental health crises, such as suicidal ideation or panic attacks. Integrating advanced crisis management algorithms or creating hybrid models that combine LLMs with real-time human oversight could significantly improve the reliability and safety of these systems in high-risk situations. Finally, given the resource-intensive nature of training large-scale LLMs, future research should explore the development of more efficient models that can be deployed in low-resource settings. The democratization of LLM technologies will be crucial for ensuring that advancements in mental health care are accessible to underserved populations, particularly in regions with limited technological infrastructure.

.

8 Appendix



Anexo 1: #Mcguffie, Kendal; Henderson-Sellers, Ann (2001): <Forty years of numerical climate modeling= Int. J. Climatol., vol. 21, 1067-1109

9 References

Adhikary, P. K., Srivastava, A., Kumar, S., Singh, S. M., Manuja, P., Gopinath, J. K., Krishnan, V., Gupta, S. K., Deb, K. S., & Chakraborty, T. (2024). Exploring the efficacy of large language models in summarizing mental health counseling sessions: Benchmark Study. *JMIR Mental Health*, *11*, e57306. https://doi.org/10.2196/57306

Anyoha, R. (2017). The History of Artificial Intelligence. Harvard University.

Baktash, J. A., & Dawodi, M. (2023). GPT-4: A Review on Advancements and Opportunities in Natural Language Processing. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2305.03195

Baxt, W. G. (1990). Use of an artificial neural network for data analysis in Clinical Decision-Making: The diagnosis of acute coronary occlusion. *Neural Computation*, *2*(4), 4803 489. https://doi.org/10.1162/neco.1990.2.4.480

Belyaeva, A., Cosentino, J., Hormozdiari, F., McLean, C. Y., & Furlotte, N. A. (2023). Multimodal LLMs for health grounded in individual-specific data. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2307.09018

Binz, M., & Schulz, E. (2023). Turning large language models into cognitive models. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2306.03917

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., . . . Amodei, D. (2020a). Language Models are Few-Shot Learners. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2005.14165

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., . . . Amodei,

D. (2020b). Language Models are Few-Shot Learners. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2005.14165

Chen, Z., Cano, A. H., Romanou, A., Bonnet, A., Matoba, K., Salvi, F., Pagliardini, M., Fan, S., Köpf, A., Mohtashami, A., Sallinen, A., Sakhaeirad, A., Swamy, V., Krawczuk, I., Bayazit, D., Marmet, A., Montariol, S., Hartley, M., Jaggi, M., & Bosselut, A. (2023). MEDITRON-70B: Scaling medical pretraining for large language models. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2311.16079

Chen, Z., Lu, Y., & Wang, W. Y. (2023). Empowering Psychotherapy with Large Language Models: Cognitive Distortion Detection through Diagnosis of Thought Prompting. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2310.07146

Cho, Y., Kim, M., Kim, S., Kwon, O., Kwon, R. D., Lee, Y., & Lim, D. (2023). Evaluating the efficacy of interactive language therapy based on LLM for High-Functioning Autistic Adolescent Psychological Counseling. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2311.09243

D9Souza, R. F., Amanullah, S., Mathew, M., & Surapaneni, K. M. (2023). Appraising the performance of ChatGPT in psychiatry using 100 clinical case vignettes. *Asian Journal of Psychiatry*, *89*, 103770. https://doi.org/10.1016/j.ajp.2023.103770

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2018a). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1810.04805

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2018b). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1810.04805

DOI: https://doi.org/10.48550/arXiv.2409.00112

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1702.08608

Dunstan, D. A., Scott, N., & Todd, A. K. (2017). Screening for anxiety and depression: Reassessing the utility of the Zung scales. BMC Psychiatry, 17, Article 329. https://doi.org/10.1186/s12888-017-1489-6

Emsley, R. (2013). Non-adherence and its consequences: understanding the nature of relapse. *World Psychiatry*, *12*(3), 2343235. https://doi.org/10.1002/wps.20067

Englhardt, Z., Ma, C., Morris, M. E., Xu, X. "., Chang, C., Qin, L., McDuff, D., Liu, X., Patel, S., & Iyer, V. (2023). From classification to clinical insights: towards analyzing and reasoning about mobile and behavioral health data with large language models. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2311.13063

Filienko, D., et al. (2024). Toward large language models as a therapeutic tool: Comparing prompting techniques to improve GPT-delivered problem-solving therapy. arXiv Preprint.

Galatzer-Levy, I. R., McDuff, D., Natarajan, V., Karthikesalingam, A., & Malgaroli, M. (2023). The capability of large language models to measure psychiatric functioning. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2308.01834

Heston, T. F. (2023). Evaluating risk progression in mental health chatbots using escalating prompts. *medRxiv (Cold Spring Harbor Laboratory)*. https://doi.org/10.1101/2023.09.10.23295321

Hirschfeld, R. M. A., Lewis, L., & Vornik, L. A. (2003). Perceptions and impact of bipolar Disorder. *The Journal of Clinical Psychiatry*, *64*(2), 1613174. https://doi.org/10.4088/jcp.v64n0209

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term memory. *Neural Computation*, *9*(8), 173531780. https://doi.org/10.1162/neco.1997.9.8.1735

https://sitn.hms.harvard.edu/flash/2017/history-artificial-intelligence/

Hua, Y., Na, H., Li, Z., Liu, F., Fang, X., Clifton, D., & Torous, J. (2024, August 21). *Applying and Evaluating large language Models in Mental Health Care: A Scoping Review of Human-Assessed Generative Tasks*. arXiv.org. https://arxiv.org/abs/2408.11288

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., De Las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., & Sayed, W. E. (2023). Mistral 7B. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2310.06825

Jin, H., Chen, S., Dilixiati, D., Jiang, Y., Wu, M., & Zhu, K. Q. (2023, November 15). *PsyEVaL: a suite of mental health related tasks for evaluating large language models*. arXiv.org. https://arxiv.org/abs/2311.09189

Jin, Q., Wang, Z., Floudas, C. S., Chen, F., Gong, C., Bracken-Clarke, D., Xue, E., Yang, Y., Sun, J., & Lu, Z. (2023). Matching Patients to Clinical Trials with Large Language Models. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2307.15051

Kerz, E., Zanwar, S., Qiao, Y., & Wiechmann, D. (2023). Toward explainable AI (XAI) for mental health detection based on language behavior. *Frontiers in Psychiatry*, *14*. https://doi.org/10.3389/fpsyt.2023.1219479

Kim, T., Bae, S., Kim, H. A., Lee, S., Hong, H., Yang, C., & Kim, Y. (2023). MindfulDiary: Harnessing large language model to support psychiatric patients9 journaling. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2310.05231

Lai, T., Shi, Y., Du, Z., Wu, J., Fu, K., Dou, Y., & Wang, Z. (2023). Supporting the Demand on Mental Health Services with AI-Based Conversational Large Language Models (LLMs). *BioMedInformatics*, *4*(1), 8333. https://doi.org/10.3390/biomedinformatics4010002

Lamichhane, B. (2023). Evaluation of CHATGPT for NLP-based Mental Health Applications. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2303.15727

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 4363444. https://doi.org/10.1038/nature14539

Lee, N. W., & Lee, N. C. (2008). A cryptographic key management solution for HIPAA Privacy/Security Regulations. IEEE Transactions on Information Technology in Biomedicine, 12(1), 34341. https://doi.org/10.1109/titb.2007.906101

Levkovich, I., & Elyoseph, Z. (2023a). Identifying depression and its determinants upon initiating treatment: ChatGPT versus primary care physicians. *Family Medicine and Community Health*, *11*(4), e002391. https://doi.org/10.1136/fmch-2023-002391

Levkovich, I., & Elyoseph, Z. (2023b). Suicide risk Assessments through the eyes of CHATGPT-3.5 versus CHATGPT-4: Vignette study. *JMIR Mental Health*, *10*, e51232. https://doi.org/10.2196/51232

Li, N. (2023). Ethical Considerations in Artificial Intelligence: A Comprehensive Disccusion from the Perspective of Computer Vision. *SHS Web of Conferences*, *179*, 04024. https://doi.org/10.1051/shsconf/202317904024

Li, Y., Wang, H., Yerebakan, H., Shinagawa, Y., & Luo, Y. (2023). Enhancing Health Data Interoperability with Large Language Models: A FHIR Study. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2310.12989

Liu, X., McDuff, D., Kovacs, G., Galatzer-Levy, I., Sunshine, J., Zhan, J., Poh, M., Liao, S., Paolo, D. A., & Patel, S. (2023). Large Language Models are Few-Shot Health Learners. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2305.15525

Loh, S. B., & Raamkumar, A. S. (2023). Harnessing large language models9 empathetic response generation capabilities for online mental health counselling support. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2310.08017

Ma, Z., Mei, Y., & Su, Z. (2023). Understanding the benefits and challenges of using large language model-based conversational agents for mental well-being support. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2307.15810

Mazumdar, H., Chakraborty, C., Sathvik, M., Mukhopadhyay, S., & Panigrahi, P. K. (2024). GPTFX: A novel GPT-3 based framework for Mental Health Detection and Explanations. *IEEE Journal of Biomedical and Health Informatics*, 138. https://doi.org/10.1109/jbhi.2023.3328350

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys*, *54*(6), 1335. https://doi.org/10.1145/3457607

Mikolov, T., Chen, K., Corrado, G. S., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1301.3781

Munmun, D. C., Pendse, S. R., & Kumar, N. (2023). Benefits and Harms of large language models in digital mental health. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2311.14693

Qi, H., Zhao, Q., Song, C., Zhai, W., Luo, D., Liu, S., Yu, Y. J., Wang, F., Zou, H., Yang, B. X., Li, J., & Fu, G. (2023). Supervised learning and large language model benchmarks on mental health datasets: Cognitive distortions and suicidal risks in Chinese social media. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2309.03564

Rehm, J., & Shield, K. D. (2019). Global burden of disease and the impact of mental and addictive disorders. *Current Psychiatry Reports*, *21*(2). https://doi.org/10.1007/s11920-019-0997-0

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, *65*(6), 3863408. https://doi.org/10.1037/h0042519

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), 5333536. https://doi.org/10.1038/323533a0

Sezgin, E., Chekeni, F., Lee, J., & Keim, S. (2023). Clinical accuracy of large language models and Google Search responses to postpartum depression questions: Cross-Sectional study. *Journal of Medical Internet Research*, *25*, e49240. https://doi.org/10.2196/49240

Shatte, A. B. R., Hutchinson, D. M., & Teague, S. J. (2019a). Machine learning in mental health: a scoping review of methods and applications. *Psychological Medicine*, *49*(09), 142631448. https://doi.org/10.1017/s0033291719000151

Shatte, A. B. R., Hutchinson, D. M., & Teague, S. J. (2019b). Machine learning in mental health: a scoping review of methods and applications. *Psychological Medicine*, *49*(09), 142631448. https://doi.org/10.1017/s0033291719000151

Staunton, C., Slokenberga, S., & Mascalzoni, D. (2019). The GDPR and the research exemption: considerations on the necessary safeguards for research biobanks. European Journal of Human Genetics, 27(8), 115931167. https://doi.org/10.1038/s41431-019-0386-5

Strachey, C. (1966). System analysis and programming. Scientific American, 215(3), 1123 127

Sunderland, M., Batterham, P., Calear, A., & Carragher, N. (2019). Self-Report Scales for Common Mental Disorders: An Overview of Current and Emerging Methods. In M. Sellbom & J. A. Suhr (Eds.), The Cambridge Handbook of Clinical Assessment and Diagnosis (pp. 2633277). chapter, Cambridge: Cambridge University Press.

Torous, J., Larsen, M. E., Depp, C., Cosco, T. D., Barnett, I., Nock, M. K., & Firth, J. (2018a). Smartphones, Sensors, and Machine Learning to Advance Real-Time Prediction and Interventions for Suicide Prevention: a Review of Current Progress and Next Steps. *Current Psychiatry Reports*, *20*(7). https://doi.org/10.1007/s11920-018-0914-y

Torous, J., Larsen, M. E., Depp, C., Cosco, T. D., Barnett, I., Nock, M. K., & Firth, J. (2018b). Smartphones, Sensors, and Machine Learning to Advance Real-Time Prediction and Interventions for Suicide Prevention: a Review of Current Progress and Next Steps. *Current Psychiatry Reports*, *20*(7). https://doi.org/10.1007/s11920-018-0914-y

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023). LLAMA: Open and Efficient Foundation Language Models. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2302.13971

United Nations. (2022). *Sustainable development goals: Ensure healthy lives and promote well-being for all at all ages*. https://www.un.org/sustainabledevelopment/health/

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1706.03762

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

Veale, M., & Binns, R. (2017). Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data & Society*, *4*(2), 205395171774353. https://doi.org/10.1177/2053951717743530

Wang, G., Yang, G., Du, Z., Fan, L., & Li, X. (2023). ClinicalGPT: Large Language Models Finetuned with Diverse Medical Data and Comprehensive Evaluation. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2306.09968

World Health Organization. (2019). *Mental health in the workplace*. https://www.who.int/teams/mental-health-and-substance-use/promotion-prevention/mental-health-in-the-workplace

Xu, X., Yao, B., Dong, Y., Gabriel, S., Yu, H., Hendler, J., Ghassemi, M., Dey, A. K., & Wang, D. (2023). Mental-LLM: Leveraging large language models for mental health prediction via online text data. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2307.14385

Yang, H., Liu, X., & Wang, C. D. (2023). FINGPT: Open-Source Financial Large Language Models. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2306.06031

Yang, K., Ji, S., Zhang, T., Xie, Q., Kuang, Z., & Ananiadou, S. (2023). Towards Interpretable Mental Health Analysis with Large Language Models. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (pp. 605636077). Association for Computational Linguistics.

Zhang, J., & Zong, C. (2015). Deep Neural Networks in Machine Translation: An Overview. *IEEE Intelligent Systems*, *30*(5), 16325. https://doi.org/10.1109/mis.2015.69