People's Democratic Republic of Algeria
الجمهورية الجزائرية الديمقراطية الشعبية
Ministry of Higher Education and Scientific Research
وزارة التعليم العالي و البحث العلمي

YAHIA FARÈS UNIVERSITY OF MEDEA
جامعة يحيى فارس-المدية
FACULTY OF SCIENCES
DEPARTMENT OF MATHEMATICS AND
COMPUTER SCIENCE



Master Dissertation

For the attainment Of the Master's Degree In Computer Science

Option : System Engineering and Web Technologies

Design And Development Of A Psychological Assessment Tool Based On LLM Agents

Realized By :

Said Seghir Akram
Safiddine Abdelfetah

supervised by :

Dr.Rahim Messaoud

Promotion : 2024/2025

Dedication

This thesis is dedicated to Our Families who supported us all along our wonderful parents, for your guidance, tough love, and belief in our abilities, for your unconditional support, unwavering patience, and the sacrifices you've made to provide with every opportunity. Your compassion and nurturing spirit have been instrumental in our success.

And to our friends, who have become an extended family, thank you for your friendship, your laughter, and for standing by us through the highs and lows. Your presence has made this journey all the more meaningful, and we are truly grateful to have you with us through the highs and lows. Your presence has made this journey all the more meaningful. We are truly grateful to have you.

Acknowledgement

First, we want to express our deepest gratitude to Almighty Allah for granting us the strength, patience, and determination to complete this work successfully. We would also like to sincerely thank our supervisor, **Dr. Rahim Mesaoud**, for his invaluable guidance and continuous support throughout our master's journey. His expertise, insightful advice, and constructive feedback have been instrumental in shaping the structure and enhancing the overall quality of this project.

We would like to extend our sincere appreciation to the members of the jury **Benaissa Azzedin , Rafaa Tahar** for the honor they bestow upon us by taking the time to read and evaluate this work.

Our dear parents are our pride and the source of our success, as they have made sacrifices to provide us with a conducive working environment and the best possible conditions. Without them, nothing would have been easy. May God protect and bless them so that we can repay them for all that they have given us. We would also like to thank our friends and family for their love and support during this process. Without their encouragement and motivation, we would not have been able to complete this journey

Abstract

This work aims to design and develop a psychological assessment system based on language models. by leveraging these latter's with advanced integration and utilization, The goal is to create an intelligent multi-agent system that can assist professionals psychotherapists and patients, This agentic tool is not only technically robust but also ethically aligned, mitigating risks such as hallucination and bias through prompt constraints and evidence-grounded response generation. utilizing techniques such as instruction tuning, domain-adaptive fine-tuning, and quantization, the platform achieves scalability and adaptability in resource-limited environments. By integrating advanced AI methodologies with mental health expertise, the system provides emotionally sensitive support to users while automating and enhancing documentation and decision-making tasks for clinicians. This work represents a significant step toward ethically grounded, context-aware, and clinically viable AI in mental healthcare applications. as it demonstrates the effectiveness of the medical assistant in enhancing healthcare delivery. Through

Keywords: Medical assistant, Psychology , Mental Health Support, therapy, Large language models, Natural language processing, Healthcare, Chatbot, Data processing, Domain-specific knowledge, AI-powered healthcare, Diagnosis

Résumé

Ce travail vise à concevoir et développer un système d'évaluation psychologique basé sur de grands modèles de langage. En exploitant ces derniers grâce à une intégration et une utilisation avancées, l'objectif est de créer un système multi-agents intelligent capable d'assister les psychothérapeutes et les patients. Cet outil agentique est non seulement techniquement robuste, mais également éthiquement conforme, atténuant les risques tels que les hallucinations et les biais grâce à des contraintes rapides et à la génération de réponses fondées sur des données probantes. Utilisant des techniques telles que le réglage des instructions, le réglage fin adaptatif au domaine et la quantification, la plateforme atteint une évolutivité et une adaptabilité dans des environnements aux ressources limitées. En intégrant des méthodologies d'IA avancées à une expertise en santé mentale, le système offre un soutien sensible aux émotions aux utilisateurs, tout en automatisant et en améliorant les tâches de documentation et de prise de décision des cliniciens. Ce travail représente une avancée significative vers une IA éthiquement fondée, sensible au contexte et cliniquement viable dans les applications de santé mentale, car il démontre l'efficacité de l'assistant médical pour améliorer la prestation de soins.

Mots-clés : Assistant médical, Psychologique, Soutien en santé mentale, Thérapie, Modèles de langage, Traitement du langage naturel, Santé, Chatbot, Traitement de données, Connaissances spécialisées, Santé alimentée par l’IA, Diagnostic.

ملخص

يهدف هذا العمل إلى تصميم وتطوير نظام تقييم نفسي قائم على نماذج لغوية واسعة النطاق. من خلال الاستفادة من هذه النماذج الأخيرة من خلال التكامل والاستخدام المتقدمين، يتمثل الهدف في إنشاء نظام ذكي متعدد الوكلاء يمكنه مساعدة المعالجين النفسيين والمرضى. هذه الأداة الوكيلة ليست قوية من الناحية التقنية فحسب، بل متوافقة أيضًا مع المعايير الأخلاقية، مما يخفف من مخاطر مثل الهلوسة والتحيز من خلال القيود السريعة وتوليد ردود مبنية على الأدلة. باستخدام تقنيات مثل ضبط التعليمات، والضبط الدقيق التكيفي للمجال، والتكميم، تحقق المنصة قابلية التوسع والتكيف في البيئات محدودة الموارد. من خلال دمج منهجيات الذكاء الاصطناعي المتقدمة مع خبرة الصحة العقلية، يوفر النظام دعمًا حساسًا عاطفيًا للمستخدمين مع أتمتة وتحسين مهام التوثيق واتخاذ القرار للأطباء. يمثل هذا العمل خطوة مهمة نحو الذكاء الاصطناعي القائم على الأخلاق، والواعي بالسياق، والقابل للتطبيق سريريًا في تطبيقات الرعاية الصحية العقلية. كما أنه يوضح فعالية المساعد الطبي في تحسين تقديم الرعاية الصحية.

الكلمات المفتاحية: مساعد طبي، علم النفس، الدعم النفسي، العلاج، النماذج اللغوية الكبيرة، معالجة اللغة الطبيعية، الرعاية الصحية، روبوت دردشة، تشخيص، ذكاء اصطناعي.

List of Acronyms

AI Artificial Intelligence

API Application Programming

CBT Cognitive Behavioral Therapy

DB Database

DL Deep Learning

EMR Electronic Medical Record

GPT Generative Pre-trained Transformer

HIPAA The Health Insurance Portability and Accountability Act

SAAS Software as a Service

LLaMA Large Language Model Meta AI

LLM Large Language Model

mHealth Mobile Health

ML Machine Learning

NLP Natural Language Processing

PHQ Patient Health Questionnaire

RAG Retrieval-Augmented Generation

RLHF Reinforcement learning with human feedback

UI User Interface

List of figures

1.1 NLP to LLMs timeline 4

1.2 The Transformer - model architecture 10

2.1 Assisted Ai Tech Use in Medicine. 29

2.2 Safe Environment between Patient and Doctor. 32

2.3 Mental Health Tracking 33

2.4 Clinical Decision support 35

3.1 Authentication and Registration Use Case Diagram –(App). 42

3.2 Guest Use Case Diagram –(App). 42

3.3 Use Case Diagram – Patient App 44

3.4 Use Case Diagram – Doctors App. 45

3.5 System Workflow. 52

3.6 LLM + RAG Workflow 53

4.1 SignUp screen 72

4.2 Login screen 73

4.3 Home screen 74

4.4 Ai & Doctor Chat screen 75

4.5 Profile screen 76

4.6 Login dashboard. 77

4.7 Overview Page. 79

4.8 Appointments Page. 80

4.9 Chat Ai Page 82

4.10 Patient Page 83

4.11 Patient Chat page. 83

4.12 Patient Details page. 85

4.13 Doctor profile page. 86

List of tables

Table1	Benefits of using fine tuning in various industries	15
Table2	LLM Agent Technical Specifications and Roles	48
Table3	Comparative Functionalities of Patient and Doctor Applications	57
Table4	Names & Tasks of Each Agent	68
Table5	Responses to the same queries from both models.	70
Table6	Comparison between multiple LLMs on the same queries	75

Contents

1 Large Language Models (LLMs) 5

1.1 Introduction 5

1.2 Concept of Large Language Models. 6

1.2.1 History. 6

1.2.2 Definition. 8

1.2.3 Core Concepts in Large Language Models. 9

1.3 Prompt Engineering in LLMs. 11

1.4 Fine-tuning LLM Models. 11

1.4.1 Introduction to Fine-tuning. 11

1.4.2 Definition. 12

1.4.3 Types of Fine Tuning. 12

1.4.4 Applications and Benefits. 14

1.5 Retrieval-Augmented Generation (RAG). 15

1.5.1 Introduction to RAG. 15

1.5.2 Definition. 15

1.5.3 Core Architecture of RAG Systems. 16

1.5.4 Advantages of RAG. 17

1.5.5 Embeddings in RAG. 18

1.5.6 Sentence Transformers. 18

1.6 The Utilities and Power of LLMs. 19

1.7 Field of LLMs. 20

1.7.1 Understanding and Generation. 20

1.7.2 LLM Agents as Chatbots and Assistants. 21

1.7.3 Education and Research. 21

1.7.4 Marketing and Content Creation. 21

1.7.5 Software Development. 22
1.8 Challenges and Considerations in LLMs. 22
1.8.1 Challenges in LLM Training. 22
1.8.2 Understanding the Outputs of LLMs. 22
1.8.3 Security Threats from Language Models. 23
1.8.4 Ethical Implications of LLMs. 23
1.8.5 Limitations in LLMs. 24
1.9 Conclusion. 25
2 AI in Mental Health and Psychotherapy. 26
2.1 Introduction. 27
2.2 General AI Technologies in Medicine. 28
2.2.1 Diagnostics. 28
2.2.2 Radiology. 28
2.2.3 Drug Discovery. 29
2.3 AI and Psychotherapy. 30
2.3.1 The Role of AI in Mental Health Support. 30
2.3.2 Ethical AI-Driven Psychotherapy. 31
2.4 Telepsychology and Remote Mental Health Platforms. 31
2.4.1 Secure Video and Chat Therapy Tools. 32
2.4.2 AI-Powered Symptom Tracking & Mood Monitoring. 32
2.4.3 Patient Engagement and Follow-up Tools. 33
2.5 LLM-Based Agents in Psychological Applications. 33
2.5.1 Therapy With LLMs. 33
2.5.2 Fine-Tuned Multi-Agent Framework for Detection and Support. 34
2.6 Clinical Decision Support for Mental Health Professionals. 35
2.7 Conclusion. 36
3 Psychological Tool System Design. 37

3.1 Introduction. 38
3.2 System Requirements. 39
3.3 System's Architecture and LLM Agents Modalities. 40
3.3.1 Interplay of Mobile and Web Applications. 40
3.3.2 Specialized LLM Agents: Design and Functionality. 46
3.3.2.1 The Formatting Agent 48
3.3.2.2 The Therapeutic Agent. 49
3.3.2.3 The Sentiment Analysis Agent. 50
3.3.2.4 The Medical Report Agent. 51
3.4 Implementation Specifics. 54
3.4.1 Patient-Facing Mobile Application. 54
3.4.2 Clinician-Centric Web App. 55
3.5 Model Training and Assessment. 56
3.5.1 Dataset Selection. 56
3.5.2 Evaluation Methods. 57
3.6 Conclusion. 59
4 Operational Framework. 61
4.1 Introduction. 62
4.2 Development Environment. 62
4.2.1 Google Colaboratory (Colab). 62
4.2.2 Python. 63
4.2.3 LangChain. 63
4.2.4 Hugging Face & Kaggle. 63
4.2.5 TypeScript. 64
4.2.6 Next.js and ShadCN UI. 64
4.2.7 React Native. 64
4.2.8 Supabase. 64

4.2.9 Git & GitHub. 64
4.2.10 Vercel. 65
4.2.11 Lucidchart. 65
4.3 Implementation Details. 66
4.3.1 Multi-Agent System Architecture. 66
4.3.2 Evaluation. 67
4.3.2.1 Fine-Tuned Models Evaluations. 67
4.3.2.2 RAG Evaluation. 69
4.3.2.3 Overall Evaluation. 71
4.4 Graphical Interface. 72
4.4.1 Mobile Application 72
4.4.2 Web App. 77
4.5 Conclusion. 87
General Conclusion. 88

References 89

General Introduction

Mental health represents a fundamental dimension of human well-being and quality of life. However, traditional psychological treatment methods—such as in-person therapy, manual assessments, and time-limited consultations—often present significant limitations in terms of accessibility, cost, scalability, and personalization. Patients may experience delayed access to care due to clinician shortages, geographic or financial constraints, or social stigma associated with seeking therapy. Moreover, conventional sessions are typically episodic, with limited continuity between encounters, making it difficult to track emotional evolution or personalize support over time. These systemic inefficiencies can undermine therapeutic adherence and outcomes, especially for individuals requiring timely and sustained mental health care.

The rapid advancement of Artificial Intelligence (AI), particularly Large Language Models (LLMs), has begun to reshape the possibilities for delivering mental health support. LLMs—deep neural networks trained on vast, diverse textual corpora—exhibit a sophisticated ability to understand, interpret, and generate human-like language. Transformer-based architectures, such as those introduced by Vaswani et al. (2017), have enabled these models to capture complex syntactic and semantic relationships across long sequences of text. LLMs such as OpenAI's GPT series, Google's BERT, and Meta's LLaMA models have demonstrated remarkable success across a wide range of natural language processing (NLP) tasks, including question answering, summarization, dialogue generation, and sentiment analysis.

The objective is to present a novel LLM-driven psychological assistant system, incorporating multiple layers of safety and enhancement techniques. These include instruction tuning, Low-Rank Adaptation (LoRA) for parameter-efficient domain adaptation, prompt engineering to control tone and intent, and grounding via RAG, which anchors model responses in verified therapeutic knowledge bases.

By combining the linguistic intelligence of LLMs with a modular, clinically informed framework, The goal of this work is to bridge the gaps in traditional psychological care. It offers a scalable, ethically aligned, and context-aware assistant capable of supporting patients with preliminary mental health evaluations, ongoing therapeutic dialogue, and clinician-ready documentation. Furthermore, the system is designed not to replace human therapists but to serve as a collaborative tool enhancing accessibility, reducing administrative burden, and enabling

continuous emotional support.

In a world facing growing mental health needs and systemic resource limitations, this work envisions an innovative path forward. It contributes not only to the field of AI in healthcare but

also to the practical deployment of intelligent, reliable, and empathetic LLM-based systems for psychological care.

evaluation methods will be developed for assessing the accuracy and reliability of the system. Strategies for developing the explainability and interpretability of the AI assistant will also be investigated for making the model's decisions more transparent. Finally, the project will study user experience design principles and adoption strategies for trust building and easy integration of the AI-based healthcare assistant into clinical workflows.

By addressing these objectives, this study aims to contribute to the development of accurate, reliable, and trustworthy LLM-based psychological assistants. These advancements have the potential to significantly enhance the mental health support and therapy outcome which will improve patient outcomes.

our thesis outline structure as following :

Chapter 1: Large Language Models (LLMs) This chapter provides a comprehensive overview of Large Language Models (LLMs), detailing how they leverage deep learning and transformer architectures for advanced natural language processing. It delves into the technical foundations of LLMs, their intricate training processes, from fine-tuning to Retrieval-Augmented Generation (RAG) methods, and explores the core features that have propelled their widespread use across diverse fields. The chapter also covers the various applications and benefits derived from LLMs, as well as the inherent challenges and ethical considerations associated with their development and deployment.

Chapter 2: AI In Mental Health And Psychology This chapter explores the transformative applications of Artificial Intelligence (AI), particularly Large Language Models (LLMs), within the medical field, with a precise focus on the psychological domain. It delves into specific use cases such as Telemedicine services, mobile health applications, and the development of medical chatbots and mental assistance tools. The chapter highlights how these AI-driven solutions contribute to improving mental health support delivery for individuals, discussing both the promising potential and the critical pitfalls and limitations of integrating AI as adjunctive care in mental health treatment and emotional well-being.

Chapter 3: Psychological Tool System Design This chapter provides a comprehensive overview of the design methodologies and approaches utilized for developing the AI-powered psychological tool based on LLMs. It covers the system's core requirements, its multi-layered architecture, and the interplay between its mobile patient-facing application and web-based clinician-centric dashboard. The chapter details the design and functionality of specialized LLM agents, including the Formatting Agent, Therapeutic Agent, Sentiment Analysis Agent, and Medical Report Agent, outlining their roles in processing user input, generating insights, and assisting in assessments and diagnoses.

Chapter 4: Operational Framework This chapter outlines the operational framework and practical implementation steps of the AI-based medical assistant designed for psychological support. It details the development environment and tools used, such as Google Colaboratory, Python with FastAPI, LangChain for LLM orchestration, Hugging Face, Kaggle, TypeScript, Next.js, React Native, and Supabase for backend services. The chapter further elaborates on the multi-agent system architecture, implementation specifics for both the mobile and web applications, and describes the rigorous evaluation methods applied to assess the utility, responsiveness, and reliability of the system within psychological health contexts.

Chapter 1 :
Large Language Models (LLMs)

1.1 Introduction

Large Language Models (LLMs) are fundamental to the field of artificial intelligence, particularly in Natural Language Processing (NLP), and have fundamentally changed how computers understand and create human languages. The dramatic growth of digital data, as well as increased computing power, have propelled LLMs to extraordinary success that has in turn shaped communication, automation, and knowledge management.

The new models of LLMs such as OpenAI's GPT-4o, released in May 2024, and Meta's Llama 4 series, released in April 2025, have changed the bar for text generation, translation, summarization, and question-answer abilities. For example, GPT-4o supports 50 languages and combines text, speech, and visual processing to create more engaging interactions with users. The capabilities of Meta's Llama 4 models are similarly multimodal, processing text and video information as well as images and audio.

The importance of LLMs does not just reside within their technical capabilities; they are changing domains such as content creation, communication, information retrieval, and knowledge extraction. Their versatility will change not only fields of work and study, but also everyday life. This chapter will address LLMs. This means that we will outline their history, basic concepts, the training and capabilities it provides, the limitations and possibilities they provide the user and how they will develop in the future. In discussing these aspects, the chapter will show what LLMs are, the impact they are having on society, to inform the background and possibilities in NLP and beyond.

1.2 Concept of Large Language Models

1.2.1 History

The evolution of large language models (LLMs) represents a significant milestone in artificial intelligence research, emerging from decades of foundational work in natural language processing (NLP). The field traces its origins to early machine translation experiments in the 1950s, most notably the Georgetown-IBM project (1954), which demonstrated the potential—and limitations—of rule-based approaches to language understanding. These systems, while conceptually straightforward, proved incapable of handling the complexity and variability of natural language, leading to waning interest following the critical ALPAC report (1966) and the onset of the first AI winter.[1]

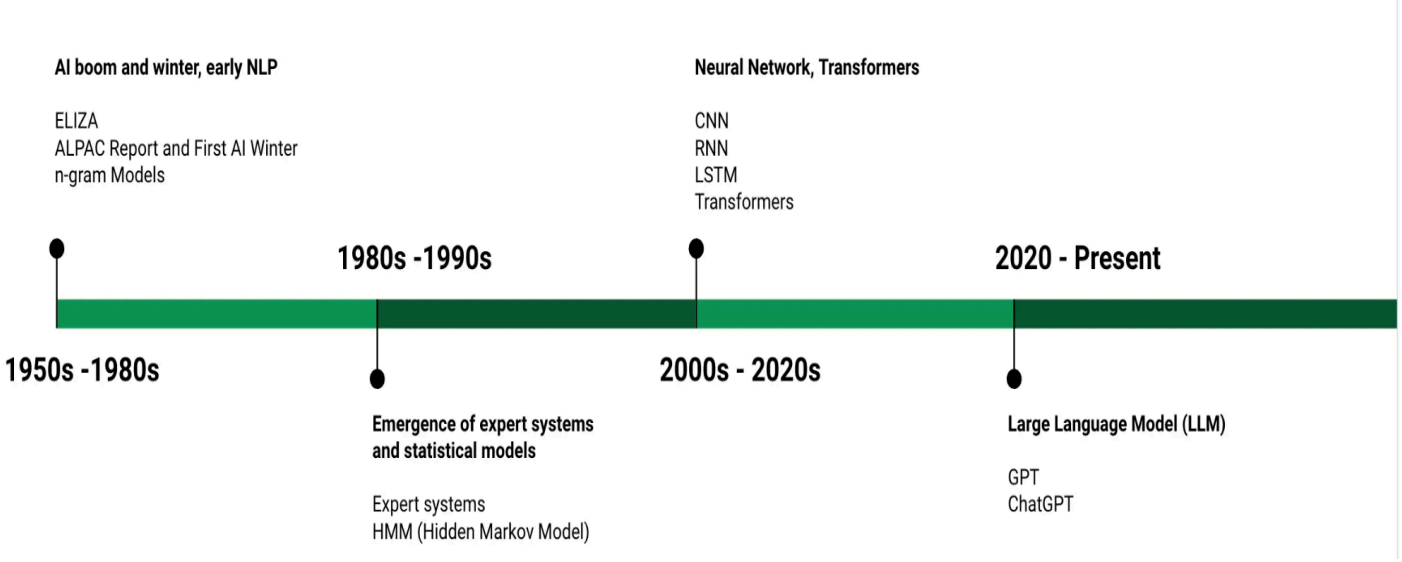


Figure 1.1: NLP to LLMs timeline

Subsequent decades saw NLP research shift toward symbolic and statistical methods. The 1970s and 1980s emphasized formal grammars and expert systems, while the 1990s introduced probabilistic models such as n-grams and Hidden Markov Models (HMMs). These techniques improved performance in specific applications—speech recognition, part-of-speech tagging—but remained constrained by their inability to capture deeper semantic relationships. A paradigm shift occurred in the 2010s with the resurgence of neural networks, enabled by increased computational power and the availability of large-scale textual data. Word embeddings (e.g., Word2Vec, GloVe) and recurrent architectures (RNNs, LSTMs) offered more nuanced representations of language, though long-range dependencies remained challenging[3].

The introduction of the Transformer architecture (Vaswani et al., 2017) marked a watershed moment. By leveraging self-attention mechanisms, Transformers enabled parallelized processing of sequential data and significantly improved context retention. This innovation laid the groundwork for modern LLMs, beginning with Google's BERT (2019), which utilized bidirectional context to achieve state-of-the-art performance across NLP benchmarks. Concurrently, OpenAI's GPT-2 demonstrated the generative potential of LLMs, albeit with acknowledged risks of misuse. The subsequent release of GPT-3 (2020) further underscored the scalability of these models, with its 175 billion parameters enabling few-shot learning and versatile task adaptation[2][4][5].

Recent advancements, including GPT-4 (2023) and Gemini 1.5 (2024), have expanded the capabilities of LLMs through multimodal processing and extended context windows. However, these developments have also intensified scrutiny around their computational demands, environmental impact, and propensity for biases or factual inaccuracies. In response, the field has increasingly prioritized efficiency-focused techniques such as retrieval-augmented generation (RAG), low-rank adaptation (LoRA), and instruction tuning—approaches that aim to maintain performance while mitigating resource use and ethical risks.

The trajectory of LLM research reflects a broader transition from rigid, rule-based systems to flexible, data-driven models capable of sophisticated language tasks. Nevertheless, key challenges persist, ensuring that ongoing work remains focused not only on scaling but also on robustness, interpretability, and real-world applicability. Future progress will likely hinge on balancing these competing priorities to realize the full potential of LLMs in both academic and practical settings.

1.2.2 Definition

Large Language Models (LLMs) are a sophisticated sub-field of artificial intelligence, precisely deep learning neural networks, often with billions of parameters, pretrained using self-supervised learning (e.g., next-word prediction or masked token filling) on extensive, diverse textual datasets. The resulting model encodes rich syntactic, semantic, and contextual information, allowing the abilities to perform tasks of understanding, interpreting, and creating human language. These models are trained on extremely large datasets that are made up of many different sources of text, including books, articles, and websites. The difference with LLMs is not just that these are very large datasets, but the size of the model itself is likewise immense, often containing billions of parameters which gives it the capacity to model the complexities and nuances of language and communication, including those found in context. The size and training of LLMs has had a profound impact on the field of natural language processing (NLP) and will allow machines to interact with humans in a more natural and meaningful way.

Notable implementations like OpenAI's GPT series and Google's BERT demonstrate the versatility of LLMs across NLP tasks, including generative (e.g., text synthesis) and discriminative (e.g., classification) applications. In specialized domains such as healthcare, these models show potential for supporting clinical workflows—ranging from symptom analysis to

preliminary therapeutic dialogue generation—though their deployment necessitates rigorous evaluation of biases, factual reliability, and computational efficiency.

A critical distinction of LLMs lies in their emergent capabilities: phenomena such as few-shot learning and instruction-following arise not from explicit programming but from the interplay of model scale, training data breadth, and architectural properties. However, these advantages are counterbalanced by challenges including high resource demands, environmental costs of training, and the need for careful alignment with human values. Current research focuses on optimizing this balance through techniques like retrieval-augmented generation (RAG) and parameter-efficient fine-tuning, which aim to preserve performance while addressing operational and ethical constraints.

1.2.3 Core Concepts In Large Language Models

Large Language Models (LLMs) operate by digesting textual corpora and discerning complex patterns, relationships, and structures that exist in natural language. These models rely upon extensive and diverse datasets, consisting of heterogeneous text, including books, scientific papers, online discussion forums, news articles, and dialogues extracted from social media, to learn how to represent and comprehend the richness of language and communication. The scale and diversity of these text corpora are important because they provide LLMs the ability to generalize across different modalities, styles, contexts, and cultures in the use of language, enabling successful performance on a wide variety of natural language, and subsequently, natural language processing (NLP) tasks [1].

LLMs learn primarily using self-supervised learning, a powerful paradigm in which the model predicts parts of the input data based on other parts of the input, lacking any human-annotated labels. For example, one model might be trained to predict masked words in a sentence (as in BERT) or predict the next word in a sequence (as in GPT). Self-supervised learning allows for rich semantic and syntactic features to be learned from unstructured text, at significantly less cost in labor for data-annotation or labeling [2]. For downstream tasks, such as sentiment analysis or medical diagnosis support, self-supervised learning is often supplemented with supervised datasets to fine-tune and specialize the model in that domain [3].

Today's LLMs are fundamentally based on the transformer architecture introduced by Vaswani et al. in the highly cited 2017 paper, "Attention is All You Need" [4]. Transformers have made a

significant impact on NLP—this change was driven in large part because of their use of a new construct, referred to as self-attention, which enables models to use all words in the sequence when assessing the importance of each word relative to each of the other words in the sequence. Recognizing that, for example, CNNs and RNNs process the sequence word by word, transformer architecture processes the entire sequence simultaneously. This has both speed implications for computation and implications for modeling potentially longer dependencies in the text [5].

The adaptability and speed of transformer-based architectures have solidified their status as the backbone of innovative models that include BERT, GPT-4, RoBERTa, PaLM, and LLaMA [6][7]. These structures have produced state-of-the-art performance on many NLP benchmarks, including question answering, summarization, translation, and dialogue generation. Furthermore, their modular structure can facilitate specialized applications in domains such as healthcare, psychology, or education, where domain-specific training may greatly enhance their accuracy and relevance[8].

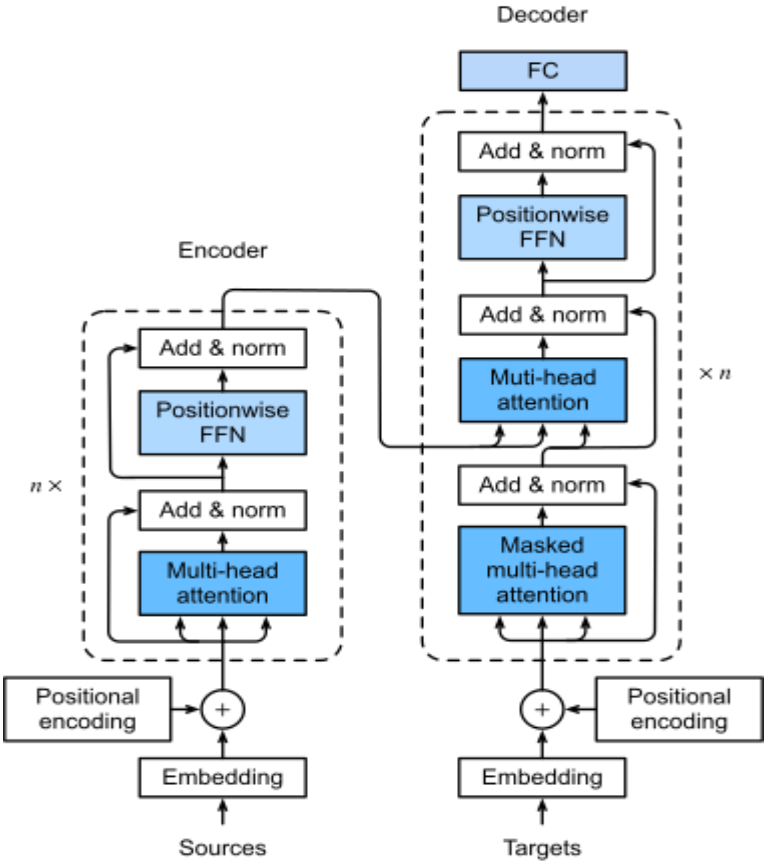


Figure 1.2: The Transformer - model architecture

The figure illustrates these components:

Encoder: Processes input text through multi-head attention and feedforward networks, generating contextual representations.

Decoder: Generates output sequences (e.g., translations or responses) using masked attention to prevent future token leakage.

This modular design enables adaptations like BERT (encoder-only) for classification or GPT (decoder-only) for generation. Current research optimizes transformers for efficiency (e.g., sparse attention) and domain specificity (e.g., retrieval-augmented decoders), balancing performance with computational constraints.

In summary, LLMs leverage extensive training data, supervised and unsupervised learning techniques, and transformer-based architectures to achieve state-of-the-art performance in natural language processing tasks.

1.3 Prompt Engineering in LLMs

Prompt engineering entails the act of creating and formulating input prompts, so LLMs produce specific responses of high quality. It is an important process to guarantee irrelevant, incoherent, or inaccurate outputs are avoided as much as possible. Examples of techniques in development of prompt engineering include:Clearly expressing the task or goal in the request.Providing contextual information or constraints for the LLM to rely on.Using formatting, keywords, or special tokens to guide the model's attention. Decomposing a complex task into smaller steps.[29] Improving the prompt by modifying it iteratively based on responses from the model.The ability to prompt engineer effectively will relate directly to the abilities of LLMs, to leverage them for all types of natural language processing and language generation tasks, and thus becoming a crucial activity for economy and should be performed for all activities with natural language processing and language generation in the real world.

For instance instead of simply asking an LLM "Explain climate change," a more effective prompt would be:

"In 100 words or fewer, explain the causes and effects of climate change to a high school student using simple language and concrete examples."

1.4 Fine-tuning LLM Models

1.4.1 Introduction To Fine-tuning

Fine-tuning has become a strong tool in the fast changing field of artificial intelligence to fit general-purpose machine learning models for particular uses. Fine-tuning lets developers take an existing, pre-trained model such as BERT, GPT, or RoBERTa and improve its performance on a particular task using a smaller, domain-relevant dataset instead of training models from scratch, which requires massive datasets, great computational resources, and substantial time commitments. Indispensable for real-world AI deployments, this approach sharpens the model's accuracy for certain use cases while preserving its general linguistic and conceptual knowledge[50].

1.4.2 Definition

Fine-tuning is the process of adjusting a pre-trained Large Language Model (LLM) to perform better on a particular task or in a specialized field. Continuing the model's training on a carefully selected, task-specific dataset updates its internal parameters (weights), therefore optimizing performance for the intended use case. Unlike building a model from scratch, fine-tuning uses the general language knowledge already ingrained in the base model to rapidly specialize with fairly tiny quantities of data.

It is Widely used in many activities including sentiment analysis, document classification, question answering, summarization, and legal contract analysis.[51] It is also customized for certain sectors including medicine (e.g., reading clinical notes), law (e.g., examining legal papers), finance (e.g., studying financial reports), and customer service (e.g., managing support tickets or chatbot inquiries). Fine-tuned models can also be optimized for various types of interaction, from structured reports and document-based responses to dynamic, conversational agents capable of multi-turn dialogue.

Fine-tuning greatly improves the accuracy, contextual relevance, and general efficacy of the model in practical uses by matching its behavior more closely with domain-specific language patterns and user expectations.

1.4.3 Types Of Fine Tuning

Fine-tuning refers to the process of further training a pre-trained large language model (LLM) on a task-specific dataset in order to specialize its behavior, adapt its responses to particular domains, or align it more closely with human intent. The fine-tuning stage builds upon the general language understanding gained during pretraining and allows models to perform with higher accuracy and relevance in specific applications. Broadly, there are three major types of fine-tuning, each with distinct methodologies, objectives, and data requirements: Supervised

Fine-Tuning (SFT), Instruction Tuning, and Reinforcement Learning from Human Feedback (RLHF).

1. Supervised Fine-Tuning (SFT)

Supervised fine-tuning is the most fundamental and widely used approach. It involves training the model on paired input-output examples, where the desired output for each input is explicitly provided by a human or a curated dataset. This approach uses traditional supervised learning methods and cross-entropy loss.

- Purpose: Adapt the model to a specific task such as question answering, sentiment classification, summarization, or domain-specific dialogue.
- Process: A dataset of prompts and target responses is used to further train the model's parameters.
- Example:
  - Input: *"Translate this sentence to French: 'Hello, how are you?'"*
  - Output: *"Bonjour, comment ça va ?"*
- Applications: Often used for building domain-specific agents or tools where precise and predictable outputs are critical.

2. Instruction Tuning

Instruction tuning is a more generalized form of supervised fine-tuning, where models are trained to follow natural language instructions rather than just task-specific inputs. The goal is to teach the model to understand and comply with a variety of user commands written in plain language.

- Purpose: Improve the model's ability to follow diverse instructions and perform many different tasks without task-specific re-training.
- Data Format: Each example consists of a user instruction and the corresponding desired response.
- Example:
  - Instruction: *"Summarize the following paragraph:"*
  - Response: *"The paragraph discusses the impact of artificial intelligence on healthcare."*
- Applications: Used in multi-task settings (e.g., FLAN-T5, InstructGPT), making the model more general-purpose and interactive.

3. Reinforcement Learning from Human Feedback (RLHF)

RLHF represents the most advanced and alignment-oriented fine-tuning method. It combines supervised learning with reinforcement learning principles, using human feedback to optimize the model's outputs toward human-preferred responses. Unlike the other methods, RLHF is not based on a fixed dataset of correct answers but instead uses comparative feedback and reward models to guide learning.

- Process:
  - Collect: Human annotators rank multiple responses generated by the model for the same prompt.
  - Train: A reward model is trained to predict human preferences.

Optimize: The LLM is fine-tuned using Proximal Policy Optimization (PPO) to maximize rewards from the reward model.

- Example:
  - Prompt: *"What's a good way to handle stress?"*
  - Human ranking favors responses that are empathetic and actionable.
- Applications: Used in models like ChatGPT, Claude, and Gemini to ensure outputs are helpful, honest, and harmless.

1.4.4 Applications and Benefits

Fine-tuning Large Language Models (LLMs) has proven to deliver tangible improvements across various industries.

Fine-tuned models offer faster deployment, taking days instead of months, a significant advantage over custom-built solutions. Additionally, they provide improved accuracy, helping performance by up to 15–60% compared to zero-shot or base models [52].

Next is a table that illustrates how various industries apply models fine tuning, highlighting the specific benefit each sector can achieve.

Industry	Use Case	key Benefit
Health Care	Medical Note Summerasation	Achieves 30% reduction in time spent on diagnosis.
Finance	Earning report Analysis	Enables real-time identification of critical financial trends and risks within earning
Customer Service	Intent Classifications	reduces misroutes and accelerates resolution. which significantly boosts customer satisfaction and operational efficiency.

table1: Benefits of using fine tuning in various industries

1.5 Retrieval-Augmented Generation (RAG)

1.5.1 Introduction to RAG

represents a major leap in the area of natural language processing through the integration of generative models with retrieval-augmented knowledge bases. RAG models generate responses based on the generation process via retrieval of documents or data from sources external to the model—not just the internal parameters. RAG models then accommodate these improvements in factual accuracy, relevance, and context by relying on text outputs grounded to both up-to-date or domain specific information. As a result, RAG is more useful for text generation than purely generative models that may be out of date, further increasing trust and reliability for tasks like question answering, summarization and domain specific chatbots [30] [31].

1.5.2 Definition

Retrieval-Augmented Generation (RAG) involves combining the generative capabilities of models like GPT (Generative Pre-trained Transformer) [32] and the accuracy and contextual grounding of retrieval-based models. RAG architectures typically consist of a retrieval module that retrieves relevant information from a large external knowledge base or document collection, which is then processed by the generative model to influence its response [33]. The collaboration between the retrieval and generative models allows RAG systems to produce fluent and coherent responses - while also being accurate and contextually relevant.

1.5.3 Core Architecture of RAG Systems

Retrieval-Augmented Generation (RAG) systems are built upon two essential components: the **retriever** and the **generator**. Together, they enable models to generate informed, accurate, and context-rich responses :

1. The Retrieval Module :

The retrieval mechanism is the foundation of RAG's ability to support responses with knowledge from the external world. It detects relevant documents or passages from a large corpus or knowledge base (e.g., Wikipedia, domain knowledge bases) and retrieves supporting knowledge for the generation process. Current retrieval methods extend beyond keyword-based searching; they use dense retrieval methods such as sentence embeddings or bi-encoders with semantic similarity with pre-trained transformer-models. Hence, RAG utilizes retrieval methods to retrieve the best possible evidence, yielding responses with factual and contextual improvements.

2. The Generation Module :

After the retrieval of relevant content, the generation module takes over. As with other components of the system, the generation module is based upon large language models (LLM), like GPT . The generation module takes the information retrieved from the retrieved context and the user input and combines the two to produce fluent and informative output.. The generator is trained to use the retrieved information properly so that the output is coherent and factually accurate [34][35]. Recent improvements in architecture and fine-tuning approaches allow the generator in RAG to adapt to a wide variety of tasks, including question answering and summarization, to produce output through the use of retrieved context as grounding.The tight integration of retrieval and generation affords RAG systems a powerful tool for natural language tasks that require both fluency and factual accuracy.

1.5.4 Advantages of RAG

Improved Coherence and Relevance in Generated Text

By dynamically integrating outside knowledge sources, Retrieval-Augmented Generation (RAG) enhances the coherence and relevance of text produced by artificial intelligence. RAG incorporates real-time retrieval mechanisms to retrieve current information prior to

generating answers, in contrast to traditional language models that only use pre-trained data and may result in responses that are out-of-date or contextually limited [36]. This hybrid approach guarantees factually correct and contextually aligned outputs.

Retrieval-augmented models are perfect for customer service, virtual assistants, and other dialogue-based applications because studies demonstrate that they maintain greater topic consistency over extended conversations [37]. A RAG-powered system, for instance, can retrieve precise product specifications or policy details from a database in a support scenario, providing accurate responses that increase user satisfaction and trust [38].

Reducing Bias with Knowledge-Guided Generation

By utilizing carefully chosen outside information, RAG also mitigates the biases present in language models. Traditional models have the potential to replicate negative or distorted viewpoints because they are trained on large, unfiltered datasets. By obtaining data from reliable, impartial sources, RAG lessens this and encourages more equitable results [39][40].
According to research, RAG's retrieval component can be adjusted to give preference to objective, reliable references, especially when discussing delicate subjects like politics or social issues [41]. For example, the model can cross-reference different points of view when discussing

controversial topics, encouraging more ethical and balanced answers [42].

Additionally, RAG's retrieval process dynamically adjusts to queries, guaranteeing generation that is both ethically sound and appropriate for the context. This adaptability is essential for creating trustworthy, objective content for a variety of applications [43][44].

1.5.5 Embeddings in RAG

A fundamental component of the RAG framework, embeddings serve as the conduit between generation and retrieval. By transforming text into dense numerical vectors, they make it possible for machine learning models to compare and analyze semantic relationships quickly. RAG can identify the most pertinent information from big datasets by using embeddings, which capture the underlying meaning of words and passages [45].

By converting a user's query into a vector representation and comparing it to knowledge sources that have been stored, embeddings in RAG enable semantic search. In order to guarantee that the generative model receives precise, context-rich data to generate high-quality responses, the system retrieves passages with closely aligned embeddings [46]. This method reduces hallucinations and factual errors while also improving response relevance and anchoring outputs in verified information.

1.5.6 Sentence Transformers

Sentence Transformers are a type of deep learning model that encodes text or phrase into high-dimensional vector representations,the model that applies sentence embeddings which means its numeric representation of sentences.

Sentence transformers main role is encoding both queries and passages in the RAG framework. These models, based on transformer architectures, are designed to generate embeddings for entire sentences, capturing the contextual nuances and semantic relationships within the text[47]. In a RAG system, sentence transformers encode queries and candidate passages into dense vectors in a shared embedding space. This shared space allows for efficient and accurate similarity comparisons, facilitating the retrieval of the most relevant passages to support the generative process[48]. By leveraging sentence transformers, RAG models can handle complex queries and retrieve contextually rich information, thereby improving the quality and relevance of the generated text[49].

1.6 The Utilities And Power Of LLMs.

Large Language Models (LLMs) have Various and remarkable proficiency in Many natural language processing (NLP) tasks. These tasks can be summarized in the followings :

a. Text Summarization

By utilizing attention mechanisms and transformer-based architecture, LLMs can efficiently summarize extensive materials into succinct and meaningful summaries. Models such as BART and PEGASUS can summarize important information from long documents, articles, or reports and provide summaries that capture clarity and relevance [9].

b. Question Answering

The performance of question answering systems has been notably improved by LLMs, which allows them to understand and respond to natural language questions with accuracy. These systems utilize a model (BERT, RoBERTa), which allows them to accurately locate and extract pertinent information from contextual information which makes them very useful in retrieving knowledge [10].

c. Text Generation

Large language models have impressive capabilities in generating well-written and contextually appropriate text in various styles and genres. For example, authorship in poetry, fiction, or journalism, GPT-3, among other models, has indicated its ability to generate human-like language with a short amount of user input, and a quick prompt is typically sufficient to generate some effective outputs [11].

d. Machine Translation

LLMs have been instrumental in improving the quality and fluency of machine translation methods. Models such as BERT and T5 have exceeded all traditional statistical approaches, developing a context-based translation of languages. Their depth of understanding of semantics and syntax has raised the bar in the field [12].

e. Code Generation

Contemporary large language models have progressed greatly in the task of generating code

and assisting programmers by producing syntactically correct and semantically relevant code in multiple programming languages.

Sophisticated models such as OpenAI's GPT-4 (through GitHub Copilot), Google's Gemini and Anthropic's Claude 3 can receive and interpret natural language prompts and generate complete functions, scripts and programs [13].

These models can support code completion and documentation, as well as assist in debugging and explaining logic, thus optimizing workflows for software developers and increasing productivity.

1.7 Field Of LLMs

Large Language Models (LLMs) have garnered significant attention due to their flexibility and potential for use across diverse fields. Their ability to understand and produce human-like language has opened up numerous possibilities across sectors.

1.7.1 Understanding and Generation

LLMs (Large Language Models) serve as the backbone for many NLP (natural language processing) applications, such as sentiment analysis, named entity recognition, and part-of-speech tagging. LLMs like BERT and GPT have been accepted in both academic and commercial roles, reliably producing state-of-the-art performance on an ever-growing multitude of NLP benchmarks[14]. Performance aside, their biggest advantage is they are able to understand the linguistic context and resolve ambiguity, which is a key component of accurately interpreting and generating language. For example, with its bidirectional training process, BERT simultaneously anticipates the context of the language before and after the word, providing a more effective approach to question answering, summary writing, etc.. GPT, through its autoregressive modeling approach, tends to excel in producing human-like fluent natural language - thinking of value-added responses as part of a conversation or helping to write a narrative.

1.7.2 LLM agents as Chatbots and Assistants

Large language models power the functionality of chatbots and virtual assistants that understand user input and respond in a natural, human-like way. GPT-based systems, including those built by OpenAI, are already used across customer service interactions, social media platforms, and smart assistants, providing custom and interactive user experiences[15]. And, the ability for LLMs to carry on and maintain conversational context across multiple turns makes conversations seem more coherent and engaging. Finally, it is now possible to fine-tune LLMs to specific industries, enhancing their realization and use in rich and specialized applications, such as mental health support, or financial advising [16].

1.7.3 Education and Research

In the area of education, Large Language Models are following up on all of the intelligent tutoring systems, automatic essay scoring, recommendation engines for modified content, etc. Educators and education researchers are also using LLMs to analyze educational materials, design individualized study plans, customized and adaptive learning platforms, etc. Automatic essay scoring systems, for instance, yield rapid and consistent feedback to students, helping

teachers scale their assessments. In addition, LLMs support the ability to provide customized educational resources for a learner, providing successive adaptive learning experiences, and achieving better learning outcomes.

1.7.4 Marketing And Content Creation

Large Language Models (LLMs) are a common tool for content creation activities, including writing blog posts, product descriptions, advertising copy, and social media. LLMs utilize user intent, along with real-time analysis of data, to assist marketers in developing tailored and engaging content that resonates with audiences[17]. There are modern tools, such as ChatGPT, Jasper, and Copy.ai, which leverage LLMs to support brands in improving their content generation processes, tracking emergent trends, and elevating their messaging for respective channels. Their capability of quickly generating high quality, contextualized text at scale supports teams in increasing productivity, advancing engagement, and fortifying their digital presence[18].

1.7.5 Software Development

Software developers are utilizing Large Language Models more and more to help with tasks like code generation, auto completion of code, bug detection, and documentation.

Tools like GPT-4 and GitHub Copilot (which is based on Codex) can translate naturally spoken or written prompting into functional code, helping to abridge software development cycles and reduce time spent generating boilerplate code-based instructions[19].

In contexts where software developers might feel overwhelmed by large codebases, LLMs can help by summarizing the function of code, adding helpful comments to code, or explaining what the code is doing from a behavioral perspective, which can be beneficial for team collaborations as well as onboarding [20].

Furthermore, LLMs can flag possible vulnerabilities or inefficiencies in code as it is being written, thus helping teams practice coding that is cleaner and more secure[21]. Putting LLMs into development environments is changing how engineers write, debug, and maintain software, and shows the growing role that AI will have in software engineering.

1.8 Challenges and Considerations in LLMs

Although Large Language Models (LLMs) show impressive abilities over a variety of tasks, widespread adoption raises a number of serious challenges and considerations that must be carefully considered.

1.8.1 Challenges in LLM Training

One of the major issues with large language models is that their training data can be biased, and the model can unintentionally propagate and exacerbate bias already present in society. Research shows that language models trained on unbalanced or biased datasets have an affinity for producing biased or discriminatory outputs, particularly in relation to demographic diversity. [22]In order to foster fairness and equity in AI systems, it is important to both critically examine the training data for bias and to develop systematic bias approaches during modeling.

1.8.2 Understanding the Outputs of LLMs

Large Language Models (LLMs), especially those built on deep neural network architectures, are often criticized for their lack of transparency and the general opaqueness of their internal reasoning. The absence of explainability and interpretability makes these intelligent agents particularly problematic in high-stakes settings such as healthcare or legal systems, where providing insight into the basis for their decisions is vital to accountability and trust. Current research aims to develop techniques that can provide transparency and interpretability for LLM outputs, which will help to varying degrees of hoped-for trustworthiness when deployed and ultimately for more responsible AI usage.[23]

1.8.3 Security Threats from Language Models

Large Language Models (LLMs) raise significant safety and security issues because they can generate harmful, misleading, or sensitive content like disinformation, hate speech, or altered stories. LLMs may also be weaponized for harm with malicious intent by bad actors to conduct phishing attacks or manipulate large campaigns of misinformation. Responding to these risks implies the construction of effective content moderation frameworks; the identification of malicious intent; and the implementation of preventative protections for the appropriate and safe use of models.

1.8.4 Ethical Implications of LLMs

The wide-ranging use of Large Language Models (LLMs) generates numerous difficult ethical dilemmas, particularly around issues of privacy, consent, and individual autonomy. The environmental ramifications from use of extensive computer power and energy required to that train LLMs warrants concerns as well. It is important to create ethical frameworks and guidelines to assess the responsible development and use of LLMs [24], and the alignment of such HCC principles with societal values and ethics.

1.8.5 Limitations in LLMs

Even though Large Language Models (LLMs) have polished skills for producing text and

undertaking other language tasks, it is important to understand that their performance is inherently limited. One of these limitations is "hallucinations," [25], when the model generates content that is factually wrong or was not a genuine attempt to comprehend the context. These limitations involve multiple contributing factors:

A. Limited Comprehension of Context: LLMs are based on statistical patterns in the training data to produce text that is grammatically correct but mostly fall short of understanding more meaningful context and essence. Therefore, their outputs can be inaccurate or nonsensical [26].

B. Restricted Knowledge Base: Even with vast amounts of training data, LLMs are still able to struggle to produce accurate and reliable information on specialized or niche topics, leading to hallucinations in the model producing responses that do not have a basis in factual knowledge.[27]

C. Difficulty Challenges and Difficulty with Ambiguity and Nuance: LLMs face challenges in dealing with language that is ambiguous or nuanced in a way that oversimplifies or misrepresents the truth, leading to fabricated or misleading information[28].

It is crucial to tackle these difficulties and concerns in order to realize the full benefit of LLMs while reducing potential risks and ensuring ethical and equi- table use.

1.9 Conclusion

Large Language Models (LLMs) have presented themselves as ground-breaking innovations in the area of natural language processing (NLP), delivering advanced capabilities that far exceed the traditional computation of language. The chapter has described in detail the fundamental principles on which they are based, their application, and indications of relevant ongoing research and development.

LLMs achieve remarkable competency in a wide variety of complex language tasks (e.g., open-ended text generation, contextual translation, abstractive summarization, question answering, or code synthesis). Their coherent, human-like language production can be attributed to their extensive parameterization, large-scale pretraining, and attention mechanisms that are aware of context.

These characteristics allow LLMs to internalize subtle syntactic and semantic patterns with high adaptability, across domains.

The landscape of applications for LLMs continues to grow rapidly. In addition to the essential NLP use cases, they are increasingly found in education, content generation, software

development, legal document review, healthcare, and scientific research. The involvement of LLMs in automating work with knowledge, generating field-specific content, and assisting in decision making demonstrates their significance across industries and potential for transforming the economy.

Looking into the future, work on LLM research will concentrate on increasing their computational efficiency, interpretability, and robustness from an ethics standpoint. Methods such as parameter-efficient fine-tuning (e.g. LoRA and adapters), instruction tuning, reinforcement learning with human feedback (RLHF), and retrieval augmented generation (RAG) have made their way into deployment of less resource-heavy and more flexible models in practice. Further, multimodal learning enables the ability of models to process and reason across text, images, audio, and structured data signals a new generation of intelligent systems that are able to more fully understand and interact in a holistic way.

Importantly, bias propagation, hallucinations, misuse of the model, and fairness continue to pose challenges that require careful mitigation. The establishment of ethical AI, alignment protocols, and continual human-in-the-loop evaluation for use in at-risk contexts is necessary for the responsible deployment of AI technologies, particularly in healthcare, law, and education.In conclusion, LLMs are reshaping the landscape of machine understanding and human-AI interaction. While substantial progress has been made thus far, continued interdisciplinary collaboration and thoughtful innovation are necessary to realize their full potential while guarding against societal and ethical challenges. As new paradigms emerge and research matures, LLMs will likely continue to lead the next decade of AI evolution.

Chapter 2:

AI In Mental Health And Psychology

2.1 Introduction

Artificial Intelligence (AI) is rapidly becoming a powerful change-agent in the healthcare world, creating innovative and improving solutions to drive diagnostic accuracy, improve treatment strategies and improve workflow in clinical care. With applications ranging from rule-based expert systems to the use of machine learning and deep learning models today, AI has evolved significantly in the field of medicine. The application of AI to healthcare signifies a multi-dimensional shift, and not only from a technological perspective, but also a philosophical perspective, in our engagement with machines in clinical settings [55].

During the last 10 years, we have seen a development of domain-specific AI systems aimed at solving certain problems across some areas of medicine. One of the more interesting and rapidly developing domains is the incorporation of AI in the realm of mental health and psychotherapy[56]. This domain has gained notable traction due to the rise in mental health matters all over the world, the limited availability of licensed mental health professionals, and the need for accessible, scalable, and personalized approaches .
The COVID-19 pandemic accelerated the digital transformation of mental healthcare and emphasized the need for telehealth platforms, virtual assistants, and intelligent monitoring services. AI-enabled solutions are already being implemented across a broad range of mental health service models— real-time mood tracking, artificial intelligence chatbots delivering

cognitive behavioral therapy (CBT), predictive analytics to detect disorders before they escalate, and virtual support groups [57].

Specifically, Large Language Models, or LLMs, such as GPT and BERT, have taken center stage in the development of new forms of therapeutic interaction. These models are capable of generating contextually relevant, empathetic responses, modeling the experience of reflective conversation, and providing journaling prompts and psychoeducational content; all of which are built upon a learning process that is primarily based on extensive language corpora [58]. These are just a few examples of how we view LLMs as promising tools for supporting the therapeutic process, especially in underserved or remote environments.
In this chapter, we will examine the expansive terrain of AI applications in medicine while focusing primarily on both mental health and psychotherapy. We will discuss the applications and uses of broader AI technologies across the healthcare system before honing in on specific applications, including psychological support, therapy augmentation, clinical decision-making, and research. [59] In this examination, we hope to shine a light on both the promise of AI and

some of the pitfalls and limitations of using AI as adjunctive care in mental health treatment and emotional well-being.

2.2 General AI Technologies in Medicine

AI technologies are innovating in many areas of medicine with sophisticated tools for diagnoses, treatments, and research. The hours/efficiency and accuracy of medical procedures will be enhanced by AI use in health and will help push developments and changes in a lot of areas. In this section, we cover a few broad areas of use AI has, in health, including radiology, diagnostics, and drug discovery.

2.2.1 Diagnostics

Diagnostic technologies using AI facilitate early disease detection, which can greatly enhance patient outcomes. Machine-learning algorithms analyze the medical data that we collect (e.g. lab test results, patient histories, imaging scans) to make inferences about disease status, identifying patterns in the data to predict disease. AI-based systems are particularly useful in identifying diseases such as cancer, diabetes, and heart disease and more precise diagnoses due to AI-assisted diagnostic tools mitigates risk of misdiagnosis and improves timely treatment.

2.2.2 Radiology

Artificial Intelligence is revolutionizing the field of radiology by streamlining image processing, helping to analyze complex images for abnormalities, and minimizing the opportunities for human error. Machine learning, and especially deep learning models, demonstrate a high degree of proficiency when parsing complex medical images, including X-rays, computed tomography (CT) scans, and magnetic resonance imaging (MRI). These tools can identify tumors, fractures, and other pathologies that may be missed by a trained human eye. AI-assisted tools allow radiologists to better manage workflow and provide more time to concentrate on complex cases.

2.2.3 Drug Discovery

The use of AI in the process of drug discovery is one of the most exciting advancements in healthcare. AI models can undertake a multitude of tasks that were once only possible either by human researchers or dominated by computer programs that could only rely on basic logic rules. AI-based models can help "predict" how various compounds are expected to behave in the body, which can lead to the identification of potential drug candidates and enable the processing of large datasets of genetic information. AI fundamentally enhances the drug discovery process by accelerating time to discovery and reproducing biological processes and their downstream effects for new treatments prior to testing in clinical trials. AI also finds potential candidates for

existing medicines, generating a shortened treatment timeline by reducing the time and cost of bringing a new therapeutic to market.

The increasing application of artificial intelligence (AI) in these sectors will improve healthcare workers' ability to make informed choices, enhance the care they provide, and contribute to innovation in medical research. There may well be further applications of AI in healthcare as technology continues to improve and evolve, and these developments will offer additional transformative changes to practice in healthcare settings.

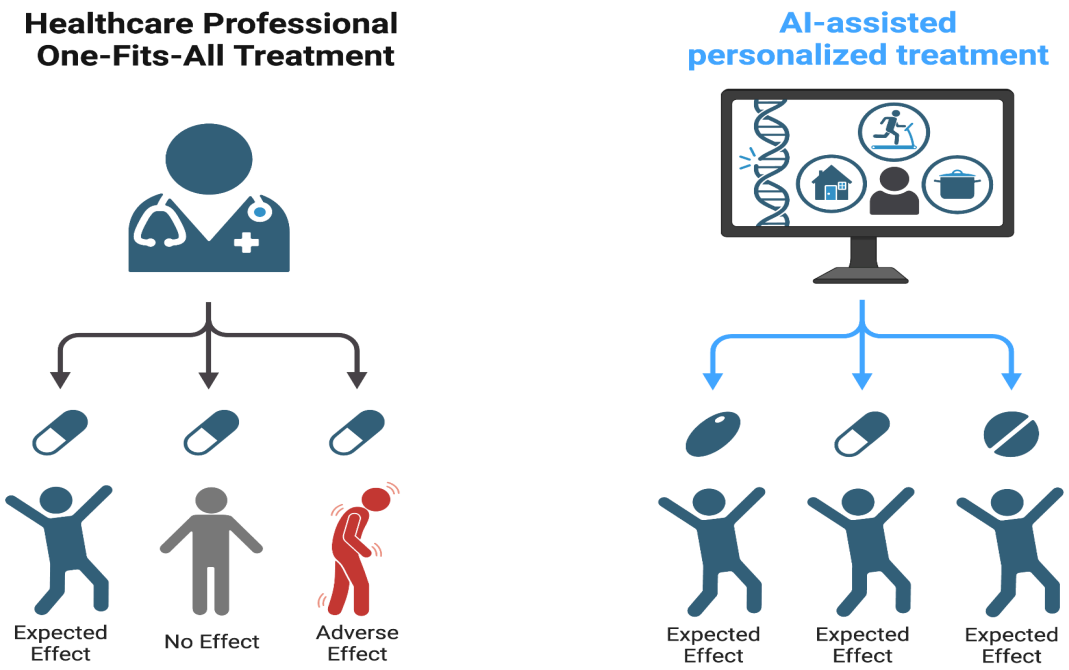


Figure 2.1: Assisted Ai Tech Use in Medicine

2.3 AI And Psychotherapy

In the last few years, the intersection of artificial intelligence (AI) and mental health has gathered considerable momentum and is catalyzing a major shift in the way psychological services are delivered, monitored, and understood.

The global mental health crisis caused, or exacerbated, in part by socioeconomic stressors, the COVID-19 pandemic, and limits to access to qualified professionals, has revealed a clear need for sustainable, scalable, new and innovative solutions.

Within this framework, AI—notably large language models (LLMs)—begins to emerge as a viable tool to extend work typically linked to psychotherapy, distribute mental health support, and contribute to new ways of psychological assessment and intervention.[60].

2.3.1 The Role of AI in Mental Health Support

Artificial Intelligence is being increasingly investigated as a disruptive technology in mental health. As the world experiences greater mental health challenges, propelled by the COVID-19 pandemic, social isolation, and economic challenges, some are suggesting AI-based solutions can be utilized to bridge the access-to-care gap. Many populations continue to experience a dearth of mental health professionals available and inadequate mental health resources. In this way, AI reminds us of its scalable utility to provide 24/7, accessible, and affordable mental health support. AI-based mental health solutions, ranging from chatbots to virtual agents, can engage users in therapeutic conversation, deliver cognitive behavioral therapy (CBT) techniques, measure mood and behavioral changes, and provide guided mindfulness practices. Training and building on large language models (LLMs), these systems typically employ fine-tuning and domain-specific adaptations to provide a deeper understanding of the users' psychological state [61].

For example, apps like Woebot and Wysa use conversational agents designed to help users regulate anxiety, depression, and stress levels. Some of these systems may detect language cues or emotional indicators, allowing these systems to tailor support to the respective person's needs.

AI is helping clinicians by automating the review of therapy session transcripts, summarizing patient progress, and providing clinical decision support based on historical data and psychological profiles.This allows therapists to spend more time on nuanced human interaction as compared to completing the repetitive tasks that AI can do without human intervention.

2.3.2 Ethical AI-Driven Psychotherapy

The application of AI in psychotherapy has the potential to provide very personalized mental health treatment. AI programs can analyze large data sets, such as electronic health records, data from wearable sensors and patterns in language use to identify early warning signs of mental distress and offer treatment proposals tailored to the individual. This early, proactive intervention is particularly important for early treatment significantly improves the outcomes for patients, such as with the clinical populations of major depressive disorder (MDD), generalized anxiety disorder (GAD) and PTSD [100][101].However, the use of AI in psychotherapy raises significant ethical issues due to the sensitive nature of the use of mental health data. Challenges around safeguarding privacy, using data for unanticipated purposes, and algorithms that contain bias have contributed to considerable concern about the use of AI in mental health applications where biased or inaccurate outputs could lead to harm [62].

Developers are incorporating privacy-preserving mechanisms, like federated learning and differential privacy, as well as ethical auditing frameworks, to address concerns about fairness, explainability and cultural competence in the AI outputs [63][64].The goal of AI in psychotherapy is not to supplant human therapists, but to advance the therapeutic process.

Hybrid models that utilize AI applications to enhance human expertise are beginning to be studied.

for overall effectiveness and user satisfaction [65]. Clinical trials are already underway to evaluate the effectiveness of AI-augmented interventions, and new regulatory frameworks are sprung up to enable responsible action in mental health care [66].

As research continues, the synergy between artificial intelligence and psychotherapy could redefine the mental health landscape making care more accessible, personalized, and continuous than ever before [67].

2.4 Telepsychology and Remote Mental Health Platforms

The swift expansion of telepsychology has revolutionized how mental health services are delivered, thereby improving access to care, most notably for individuals in remote or lack thereof service locations. AI technologies are becoming a larger part of this platform as therapy becomes more efficient, and optimized for ongoing assessments and delivery specific to the individual.

2.4.1 Secure Video and Chat Therapy Tools

Secure video and messaging platforms that are HIPAA compliant, now allow for delivered therapy sessions in real-time between therapists and the patient, regardless of location. These platforms diminish barriers to access and increase our flexibility in the delivery of mental health care. AI is often used behind the scenes to assist with transcription, sentiment analysis, and documentation of the therapist, making the clinical workflow more seamless.

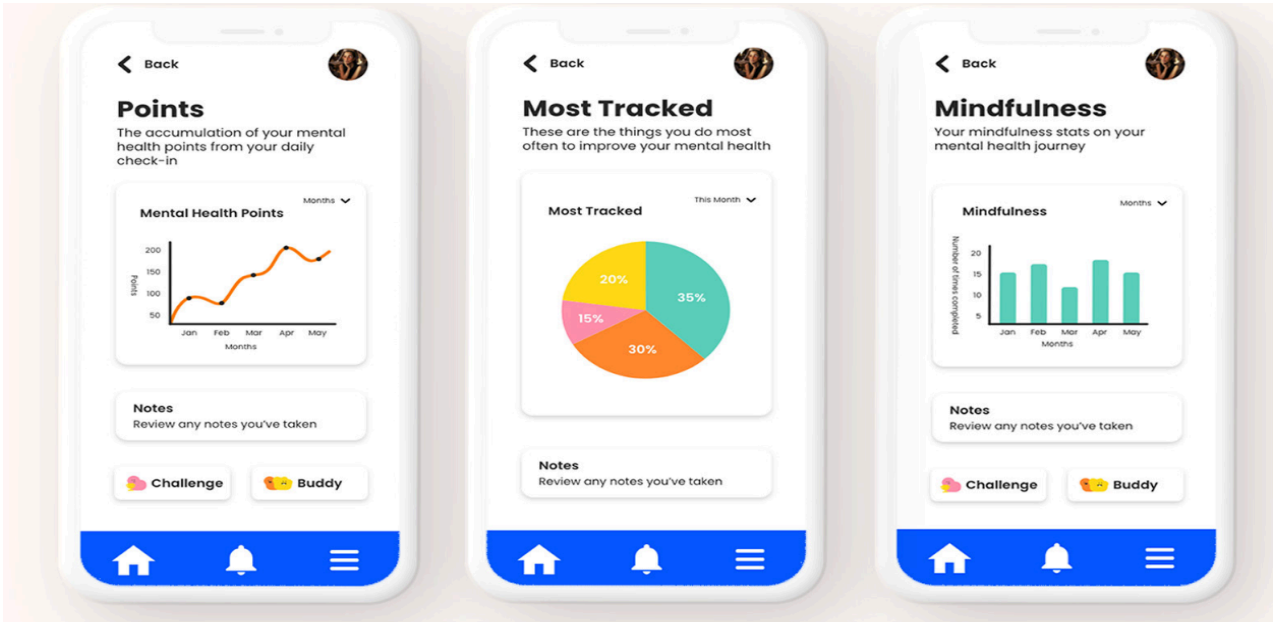


figure 2.2 : Safe Environment between Patient and Doctor

2.4.2 AI-Powered Symptom Tracking & Mood Monitoring

Mobile apps and wearables powered by AI-centric technologies can facilitate ongoing collection of user-reported symptoms, behaviors, sleep, or even tone of voice or facial expressions. The collected data feeds into models that predict mood patterns over time and can help therapists detect small shifts in mental health and take action earlier. Machine learning models are being used more and more to help produce unique insights and recommendations about therapeutic care[68].

Figure 2.3 : Mental Health Tracking



2.4.3 Patient Engagement and Follow-up Tools

Promoting patient engagement outside of therapy sessions is an important consideration for behaviors to change sustainably. With AI-enhanced tools, reminders can be automated or users can receive self-care exercises daily or content tailored to engagement or mood users had at prior sessions. Chatbots or virtual agents can also offer 24/7 support when an individual is engaging in CBT techniques, journaling, or relaxation[69].

2.5 LLM-Based Agents in Psychological Applications

2.5.1 Therapy With LLMs

There has been a growing adaptation of Large Language Models (LLMs) to simulate therapeutic-type environments that provide structured conversations that could also respond to emotional content. Often, these models are behind many psychological tools—including journaling bots, reflective dialogues agents, and Cognitive Behavioral Therapy (CBT) pre-emptive agents—when they take cues from patterns in user language, prompting the user to assess cognitive distortions, reframe related thoughts, and increase emotional regulation—which is done without the position of or replacing the therapist.These types of systems can benefit from the contextualized fluency of the LLMs to produce the type of response the user would find to be both supportive and suitable to the context in which they are situated. These models are also being increasingly integrated with reinforcement learning techniques that allow the model to adapt to user tone and intent, which may better support long-term user engagement and relevance.

2.5.2 Fine-Tuned Multi-Agent Framework for Detection and Support

The system adopts a modular multi-agent architecture, having multiple language models specialized for psychological support. Each agent operates independently yet communicates within a structured pipeline to provide end-to-end mental health assistance.

The first stage involves a context-aware conversational agent that handles the interpretation of user input and maintains coherence in multi-turn dialogues. Fine-tuned on instruction-based datasets and integrated with a retrieval-augmented generation (RAG) mechanism, this component ensures that responses are informed by relevant knowledge while remaining empathetic and context-sensitive.

Next is a lightweight classification agent, optimized for early symptom detection. This model is fine-tuned on domain-specific mental health data to recognize patterns associated with common psychological conditions such as anxiety and depression. Its computational efficiency makes it the most reliable and suitable for real-time use, including deployment on local or resource-constrained environments.

The final stage needs a higher-capacity generative model capable of performing advanced reasoning tasks. This agent generates comprehensive summaries, simulates pseudo-diagnostic insights, and offers personalized support strategies aligned with therapeutic communication practices. Its architecture supports the processing of nuanced inputs and the generation of structured, informative outputs.

This fine-tuned multi-agent system allows for scalable and modular deployment, supporting a wide range of mental health scenarios while maintaining a balance between performance, interpretability, and responsiveness.

2.6 Clinical Decision Support for Psychotherapists

Artificial Intelligence (AI) technologies serve an important role in supporting mental health professionals with their decision-making processes in clinical settings. These systems are developed to support therapists and psychologists by examining massive amounts of patient information, identifying patterns, and suggesting therapy options. One of the more significant use cases is the case-based analysis, where AI can examine a patient’s history, symptoms, and level of progress and make more impactful insight with this information, which may not occur as quickly if at all. AI from heuristic-based early warning systems can apply to predict changes in treatment/developmental trajectories enabling better clinical practice interventions. Adding AI tools during the clinical process changes the scope of mental health interventions' effectiveness by supporting a better
decision-making process for the clinicians so that more bio-psychosocial care can be made possible.

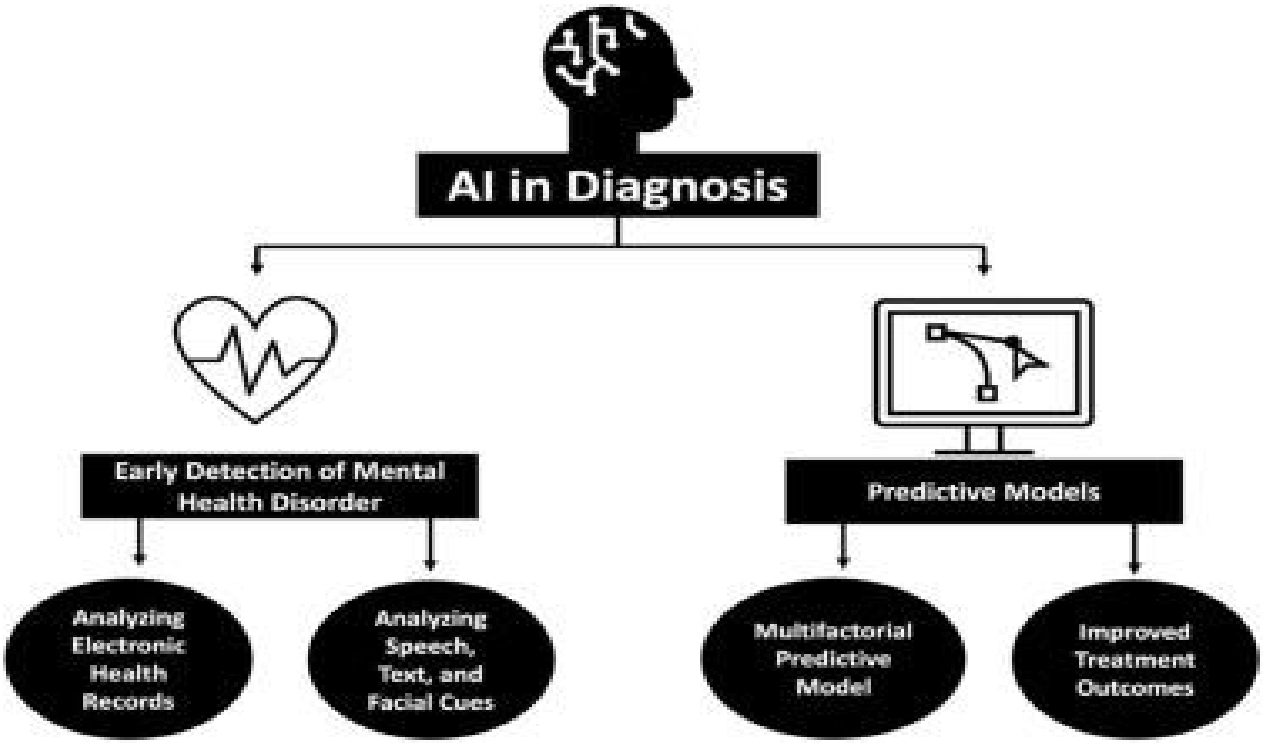


Figure 2.4: Clinical Decision support

2.7 Conclusion

As AI advances, there is a tremendous worth of its future application in mental health care. We will likely see AI integrated further into personalized treatments, stronger early detection of mental health disorders, and increased access to mental health services in underrepresented populations. The presence of AI will expand its role in context-based and case-oriented mental health service professionals in case formulation, treatment suggestions, and collaborating with the mental health research community, thus, increasing efficiency and effectiveness. The auspicious future of AI in mental health indicates a large change, where technology complements the human conduct of practice and thus, improves patient presentations with mental health; while broadening the equity of mental health services globally. Furthermore, the continual improvements in these technologies will expand potential opportunities for the delivery of mental health care in ways we have not yet imagined.

Chapter 3 :

Psychological Tool System Design

3.1 Introduction

Speed and precision of psychological assessments are essential for successful treatment and long-term results, but those are often impeded by symptom complexity as well as individual differences

Presenting a diagnosis is dependent on the qualifications and experience of the mental health professional, who can be subjective to biases, clinical exposures, and training experiences that may all result in delayed treatment, missed diagnoses, or misdiagnoses. In recognition of these challenges, we are recommending the use of AI large language models (LLMs) to serve as a mental health assistant. LLMs are AI models that have been trained on diverse sources of textual content, which makes conversational engagement with patients insightful, contextually aware, and able to interpret vague details provided by presents when discussing their symptoms.

LLMs do not replace human professionals but can augment human clinical decision making, supporting the human experience and access to care and providing a large volume of clinical knowledge, which can limit and reduce diagnostic errors, increase personalized care, and broaden access to care. In all situations, the goal is to improve clinical work and efficiency in mental health care rather than replace the human professional.

complements the human conduct of practice and thus, improves patient presentations with mental health; while broadening the equity of mental health services globally. Furthermore, the continual improvements in these technologies will expand potential opportunities for the delivery of mental health care in ways we have not yet imagined.

3.2 System Requirements

Our work is to design and develop an assessment system for psychotherapy and mental health support which will be offered to psycho-therapists , mental health doctors and professionals . as well as patients.

The proposed system functions as a comprehensive AI-powered psychological assistant and assessment tool, meticulously designed to support mental health evaluation and user well-being through natural language interaction. Its architecture is bifurcated into two primary applications: a mobile application specifically developed for patient interaction and a web-based dashboard exclusively tailored for mental health professionals.[1] This dual-application structure is a strategic architectural decision that optimizes the user experience and utility for distinct user groups while maintaining a unified backend. While both applications are engineered to complement each other, enabling seamless

communication and data flow between patients and their assigned doctors, each is also fully capable of delivering a standalone user experience. Their combined use, however, significantly enhances the overall system efficacy by allowing their agents to act as complementary elements.

Central to this system are four specialized LLM agents, each with a distinct function: a prompt formatting agent, a therapeutic response generation agent, a sentiment analysis agent, and a medical report generation agent. These agents collectively interpret user input, generate mental health insights, and assist in reflective self-assessment or professional diagnoses.[1] The overarching objective of this system remains consistent: to enhance clinical work and efficiency in mental health care through AI augmentation, ensuring that human professionals remain central to the treatment process.[1] This nuanced approach to system design reflects a deep understanding of the divergent requirements and ethical considerations for patients (e.g., direct empathetic support, self-help tools) versus clinicians (e.g., comprehensive patient management, detailed diagnostic reports), thereby maximizing utility, safety, and adherence to professional standards for each user type.

The system focuses heavily on utilizing LLMs as the support needed for the doctors to clear up any ambiguity or biases . having that LLMs once trained on sentiment detections and psycho-counseling conversations can provide significant help in both doctors report and diagnoses in addition to contributing on giving more context to the doctor by providing empathetic responses which allows the patients to feel more safe sharing some crucial sentiments or experience that can substantially boost the treatment .

These LLMs require meticulously structured prompts to ensure the efficiency of the provided support. especially in a sensitive field like mental health , where any misunderstanding or confusion can crucially astray the conversation’s purpose and slows or might even affect the treatment process negatively. As a result we needed to guarantee that these prompts(queries / messages) coming from both the Doctors and the patients are engineered to guide the model in accurately capturing the nuances of emotions and expression for an effective diagnosis. Understanding the tone and emotional cues coming from the queries / prompts allow the LLMs to dynamically understand and adjust its own responses , guiding the conversation in the most therapeutic direction. This ability to adapt its communication style based on patient sentiment is key to providing tailored

assistance and enhancing the overall treatment journey.

This engineering fine-tunes the model's empathetic output . Additionally it ensures ethical boundaries and patient safety are paramount. Ultimately, sophisticated prompt design is what transforms a generic AI into a truly specialized and effective therapeutic aid.

3.3 System’s architecture and LLM agents modalities

The Structure of our developed AI psychological assistant and assessment system, which is designed to support mental health evaluation and user well-being through natural language interaction. The system integrates Large Language Models **(LLMs)** to interpret user input, generate mental health insights, and assist in reflective self-assessment or professional diagnoses.

3.3.1 Interplay of Mobile and Web Applications

The system's overall architecture is conceptualized as an integrated AI psychological assistant and assessment system, designed to support mental health evaluation and user well-being through natural language interaction.[1] This architecture is fundamentally built upon a dual-application structure: a mobile application dedicated to patients and a web-based application exclusively for clinicians.[1] These two applications are engineered to complement each other, facilitating seamless communication and data flow between patients and their assigned doctors, while each also retains the capability to deliver a standalone user experience. When utilized in conjunction, the system and its integrated agents act as complementary elements that enhance and complete one another.[1]

The system incorporates an authentication and registration process, emphasizing a role-based access control mechanism.[1] Users can register either manually or via Google, subsequently selecting their user type as either 'Client' or 'Doctor' to complete their profile. Both processes lead in navigation to the main application page.[1] This reliance on role-based access control and distinct user interfaces for patients and doctors is a foundational principle for both security and usability. It ensures that sensitive medical data and advanced diagnostic tools are accessible only to authorized professionals, while patients receive tailored, safe interactions, thereby reinforcing the ethical framework .

Detailed use case diagrams further illustrate the system's functionalities and user interactions. The first and foremost is the Auth Use Case Diagram (figure 3.1) ,The Guest Use Case Diagram (Figure 3.2) shows how unauthenticated users can access the AI

Chatbot, offering text and optional vocal messaging. The Patient App Use Case Diagram (Figure 3.3) outlines interactions for authenticated patients and parents, including access to meeting dates, history, profile, doctors, and the Chatbot. Both voice and chat modes for the Chatbot lead to diagnosis generation and report downloads. Within the Patient App, users can also manage assignments and message doctors directly. The Doctors App Use Case Diagram (Figure 3.4) details the clinician's workflow, including access to Assistant, Patients, and Profile pages, enabling file and message management, diagnosis generation, report downloads, patient chat, history review, and meeting management.

The main architecture of the system is as specified by the following figures :

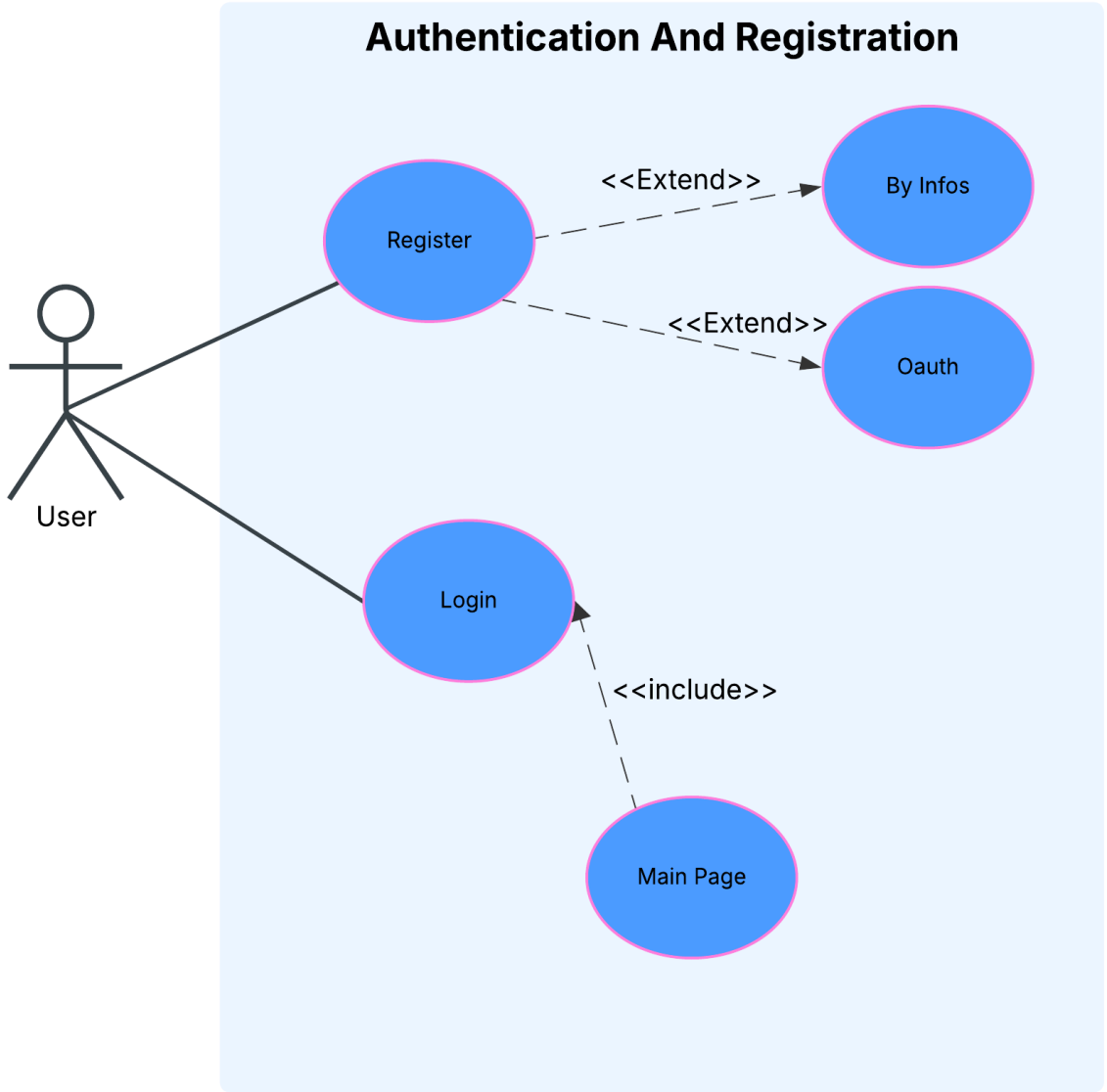


Figure 3.1: Authentication and Registration Use Case Diagram –(App)

of mere automation; it acts as an intelligent clinical assistant that supports more reliable

diagnostic outcomes by unifying emotionally nuanced user interactions and structured medical data [96][97].

This use case diagram represents the authentication and registration process within the system. The primary actor, the User, can either Register or Login. The registration process extends to two methods: By Info (manual entry) and Google (third-party login). After registration, the user selects a User Type, either Client or Doctor, to complete their profile. Both registration and login processes include navigation to the main page. The diagram illustrates a role-based access approach with flexible authentication methods.

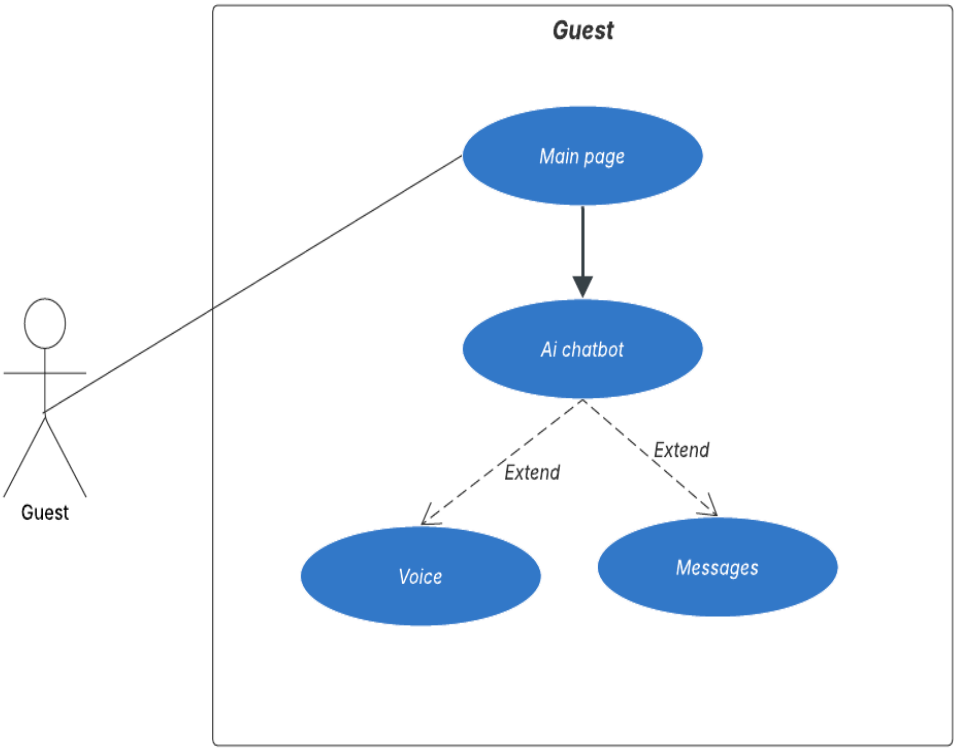


Figure 3.2: Guest Use Case Diagram –(App)

The diagram shows how a Guest interacts with the system's functionalities. The user interaction begins on the Main page as the entry point for the guest.

The Main page provides access to the AI Chatbot, which is the primary function of the system. The AI Chatbot can provide messaging options through text and vocal options.

The use case for the AI Chatbot is extended by two optional use cases titled Voice. These use case extensions provide optional features that can subsequently facilitate chatbot interaction, either through voice input or output functionality. The extend relationships show that the voice features are optional features that rely on the base use case of the Chatbot

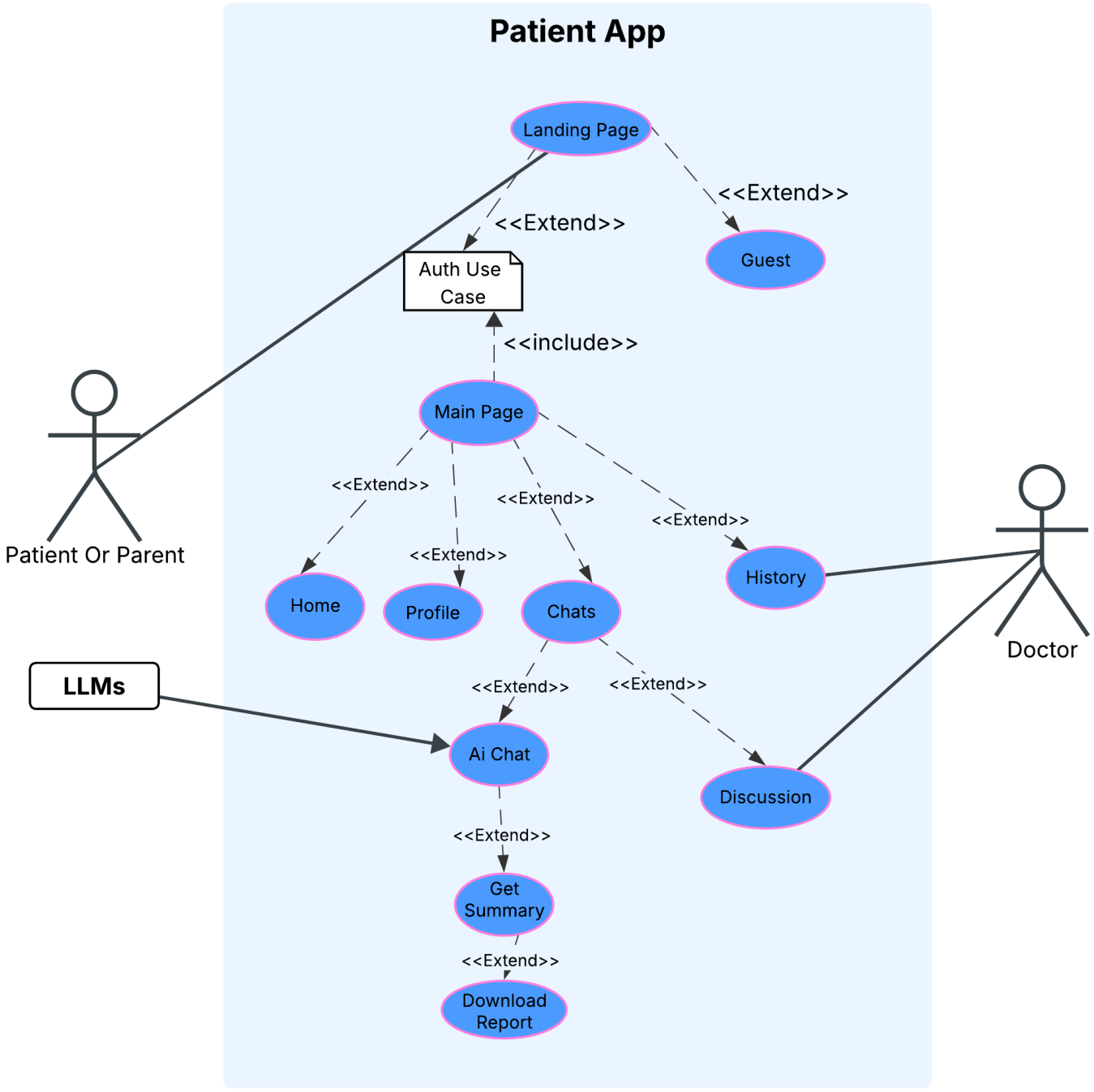


Figure 3.3: Use Case Diagram – Patient App

This use case diagram describes the interactions between a patient or parent and the Patient App system. Once the patient has accessed the system they start at the Main page and can either Authenticate through Auth or use the system as a Guest, which triggers Guest use cases.

If the user decides to Authenticate, they are brought to the Home page that is an access point for features such as meeting dates, history, profile, Doctors and Chatbot.

The Chatbot has two interaction modes : the user can choose to interact through Voice or Chat but both modes will direct them to the use case Get a diagnosis and then allow them to

Download rapport.

Inside the Doctor's page, the user can see their Assignments where they can also Accept, Edit, or Delete to manage the assignments, and it also has the ability to message the doctors separately.

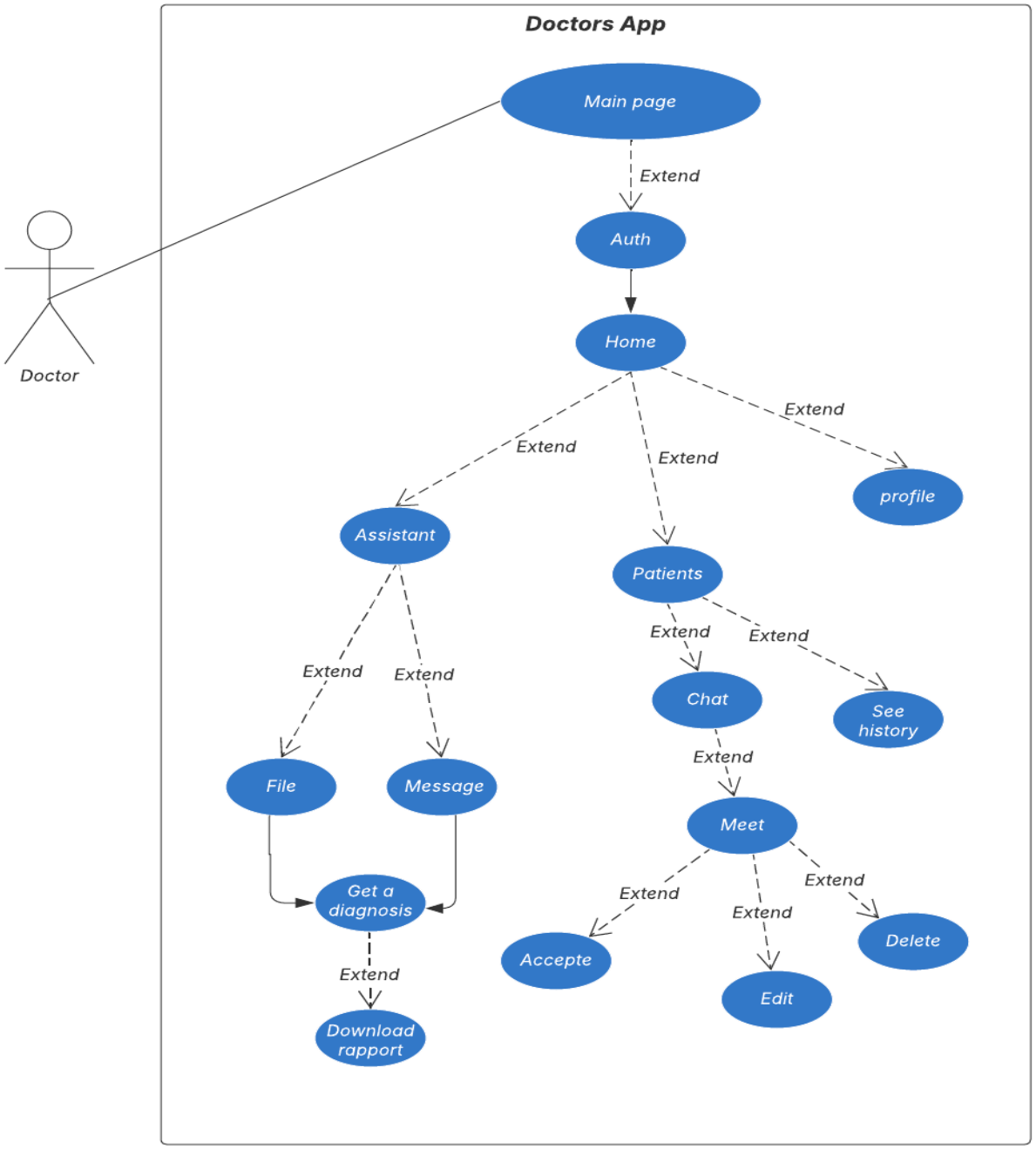


Figure 3.4: Use Case Diagram – Doctors App

This diagram illustrates the doctor's role with the Doctors App system, starting with the Main page, from which the doctor authenticates via Auth and goes to the Home page.

Once on the Home page, the doctor may view Pages such as Assistant, Patients, or Profile.

The doctor can manage Files and Messages in the Assistant Page, both of which lead to the Get a diagnosis function, as well as to download rapport.

Once on the Patients Page the doctor can Chat with patients, See history of the patient, or Meet a patient. When selecting the Meet, the doctor can Accept, Edit, or Delete the Meet.

3.3.2 Specialized LLM Agents: Design and Functionality

The system’s core intelligence is driven by four distinct Large Language Model (LLM) agents, each meticulously designed for a specific function within the mental health support framework. These agents collectively interpret user input, generate insights, and assist in assessment and diagnosis.

Agent Name	Primary Function	Key Technical Mechanisms/Concepts	Contribution to System
Prompt Formatting Agent	Standardize and optimize user input (queries/messages) from both patients and doctors.	Rule-based prompt engineering, Instruction-based prompting, Role-defining prompts, Domain-specific prompts.	Ensures LLM efficacy, accuracy, and safety by guiding models to capture emotional nuances and expressions; prevents misinterpretation in sensitive contexts.
Therapeutic Response Agent	Generate empathetic, context-aware, and informed therapeutic dialogue for patients.	Fine-tuning on psycho-counseling datasets, Retrieval-Augmented Generation (RAG) for external knowledge integration, Dynamic memory systems, Parameter-efficient fine-tuning (e.g., LoRA).	Fosters patient trust and engagement, encourages open expression, provides personalized and clinically grounded advice, enhances overall treatment journey.
Sentiment Analysis Agent	Detect and interpret the emotional tone and cues embedded in user inputs (text and potentially voice).	Natural Language Processing (NLP) for emotional recognition, Multi-modal AI integration, External APIs for real-time analytics.	Enables adaptive conversation flow by allowing LLM to dynamically adjust its communication style, tailoring assistance to patient sentiment, and boosting treatment efficacy.

Medical Report Generation Agent	Produce comprehensive assessment diagnoses and detailed medical/therapeutic reports for clinicians.	Natural Language Processing (NLP), Machine Learning (ML), Generative AI, Data synthesis from chat transcripts and clinician notes.	Supports clinician diagnosis and workflow efficiency, enhances final reports by integrating multiple data sources, reduces human error and biases, ensures consistency and compliance.

Table 2: LLM Agent Technical Specifications and Roles

3.3.2.1. The Formatting Agent

At the core of our intelligent system design lies a dual-function preprocessing and prompt engineering mechanism that ensures both the clarity of input and the therapeutic efficacy of output. The formatting agent, built upon a proficient grammar correction model, works on cleaning user inputs through grammatical correction and structural normalization. This preprocessing step enhances input coherence and prepares the text for more effective interpretation by downstream agents. In addition to fixing grammatical errors, the formatting agent removes noise and unwanted characters, eliminates redundant white spaces, standardizes case formatting, and imposes consistent structural patterns on the incoming queries. These foundational corrections are essential not only for linguistic clarity but also for semantic precision, which plays a critical role in the next stage of prompt engineering.

Prompt engineering, defined as the art and science of structuring inputs to guide LLM responses, plays an especially critical role in mental health applications, where the stakes of miscommunication are high and emotional sensitivity is paramount [78][79][81]. The system leverages a prompt formatting agent that translates pre-processed queries into carefully crafted instructions designed to capture subtle emotional cues, psychological nuances, and clinical context embedded in both patient and therapist inputs [80][83]. This stage employs several techniques, including instruction-based prompts that define output format and tone, elaborated prompts enriched with contextual cues, role-defining prompts that shape model identity and behavior (e.g., acting as a therapist), and domain-specific prompts tailored to align with specialized psychological knowledge [85]. This meticulous engineering process not only enhances the empathy and relevance of generated responses but also integrates ethical safeguards, ensuring that the LLM operates within defined safety parameters [79][81]. By constraining the LLM's behavior through explicitly structured prompts, the system mitigates risks of bias, misinformation, or inappropriate replies—ultimately transforming the model into a

reliable and ethically grounded therapeutic instrument [78][80][84].

3.3.2.2 The Therapeutic Agent

The Therapeutic Response Agent forms the emotional and cognitive core of the patient-facing system, designed to deliver empathetic, contextually aware, and clinically informed replies to user inputs [79][81]. After the user's input is grammatically refined by the formatting agent, it is passed to the Therapeutic Agent, which operates on an optimized Instruct model integrated with a Retrieval-Augmented Generation (RAG) mechanism via LangChain. This architecture is engineered to support meaningful mental health interactions by embedding the query, retrieving relevant therapeutic information from a domain-specific knowledge base, and crafting a response that is both emotionally resonant and grounded in psychological evidence [86][90].

The response generation process unfolds in three stages: sentence embedding, vector storage, and semantic retrieval. First, the user's text is converted into an embedding—a numerical vector that captures the semantic essence of the message. This vector is then stored and compared within a specialized vector database. Finally, based on similarity matching, the system retrieves contextually relevant therapeutic content, which informs the construction of a sensitive, personalized response tailored to the user's emotional state and psychological needs. This process, driven by RAG, mitigates the hallucinations often seen in general-purpose LLMs by anchoring the response in validated external knowledge [87][90][96].

To further enrich its capabilities, the agent undergoes domain-specific fine-tuning using psychotherapy-focused datasets and employs parameter-efficient tuning strategies such as Low-Rank Adaptation (LoRA), allowing the LLM to operate effectively within healthcare constraints while requiring relatively few high-quality data samples [84][87][94]. These enhancements ensure that the system understands clinical nuances, maintains ethical integrity, and remains adaptable across a wide range of therapeutic interactions. The inclusion of a dynamic memory system further contributes to conversational continuity, enabling the agent to personalize its support across multiple sessions and build therapeutic rapport over time [91]. Altogether, the synergy of fine-tuning and RAG creates a robust and ethically aligned foundation for delivering high-quality, human-like mental health support through AI [86][88][90][96].

3.3.2.3. The Sentiment Analysis Agent

The Sentiment Analysis Agent is a core component of the system's adaptive intelligence, responsible for interpreting the emotional tone and cues embedded in user inputs—whether through text or, in future iterations, potentially voice-based interactions [79]. This emotional understanding is fundamental to therapeutic success, as it allows the LLM to dynamically adjust its tone, language, and content to align with the user's emotional state, thus guiding the conversation in a more supportive and therapeutic direction [83]. To achieve this, the system employs a fine-tuned model trained specifically for sentiment classification in mental health contexts. This model processes each user input, particularly at the end of a session, to generate a comprehensive sentiment profile. This emotional summary is then passed to downstream agents, particularly for documentation and reporting, ensuring that clinicians receive a rich and emotionally contextualized overview of the interaction.

The sentiment analysis process not only enables tailored responses but also fosters a psychologically safe environment where users feel more comfortable disclosing personal and sensitive information [81][82]. By moving beyond static, pre-programmed chatbot behavior, the system adopts a more nuanced, human-like therapeutic dialogue driven by real-time emotional insights [80][83]. Additionally, multi-modal capabilities are envisioned—integrating voice and textual sentiment analysis to capture subtler emotional cues and elevate the model's understanding of user needs [91]. This adaptive, emotion-aware interaction is critical for building rapport and trust—cornerstones of effective psychotherapy—and directly contributes to improved patient outcomes by encouraging openness, ensuring emotional resonance, and promoting deeper therapeutic engagement [81][83][96].

3.3.2.4. The Report Agent

The Medical Report Generation Agent serves as a vital bridge between AI-driven patient interaction and clinical decision-making, with its core function being the automated production of comprehensive psychological assessments, diagnoses, and structured therapeutic reports tailored for clinicians [78]. Utilizing diverse data inputs—including patient-therapist dialogue, sentiment analysis results, and optional clinician notes—the agent synthesizes a holistic and evidence-based overview of the patient's condition [82]. This multi-source integration enables the generation of both detailed medical reports for clinical use and concise interaction summaries for non-clinical purposes, such as personalized feedback, emotional validation, and

lifestyle recommendations.

Technically, the agent is built upon a quantized large language model, optimized for resource efficiency and clinical robustness. Its architecture employs Natural Language Processing (NLP), Artificial Intelligence (AI), and Machine Learning (ML) techniques to convert raw inputs—ranging from chat transcripts and diagnostic test results to free-form clinical notes—into well-structured, standardized, and easily interpretable documents [95]. The use of AI-driven report synthesis not only enhances diagnostic clarity but also offers a layer of objectivity by mitigating potential biases in human interpretation and cross-referencing multiple information sources [81][95].

Beyond improving the accuracy and consistency of clinical documentation, the agent addresses significant pain points in healthcare workflows. By automating time-consuming documentation processes, it helps reduce clinician burnout, increases operational efficiency, and improves regulatory compliance through standardized formatting [93][95][97]. This enables mental health professionals to dedicate more time to direct patient care. Critically, the agent transcends the role of mere automation; it acts as an intelligent clinical assistant that supports more reliable diagnostic outcomes by unifying emotionally nuanced user interactions and structured medical data [96][97].

of mere automation; it acts as an intelligent clinical assistant that supports more reliable diagnostic outcomes by unifying emotionally nuanced user interactions and structured medical data [96][97].

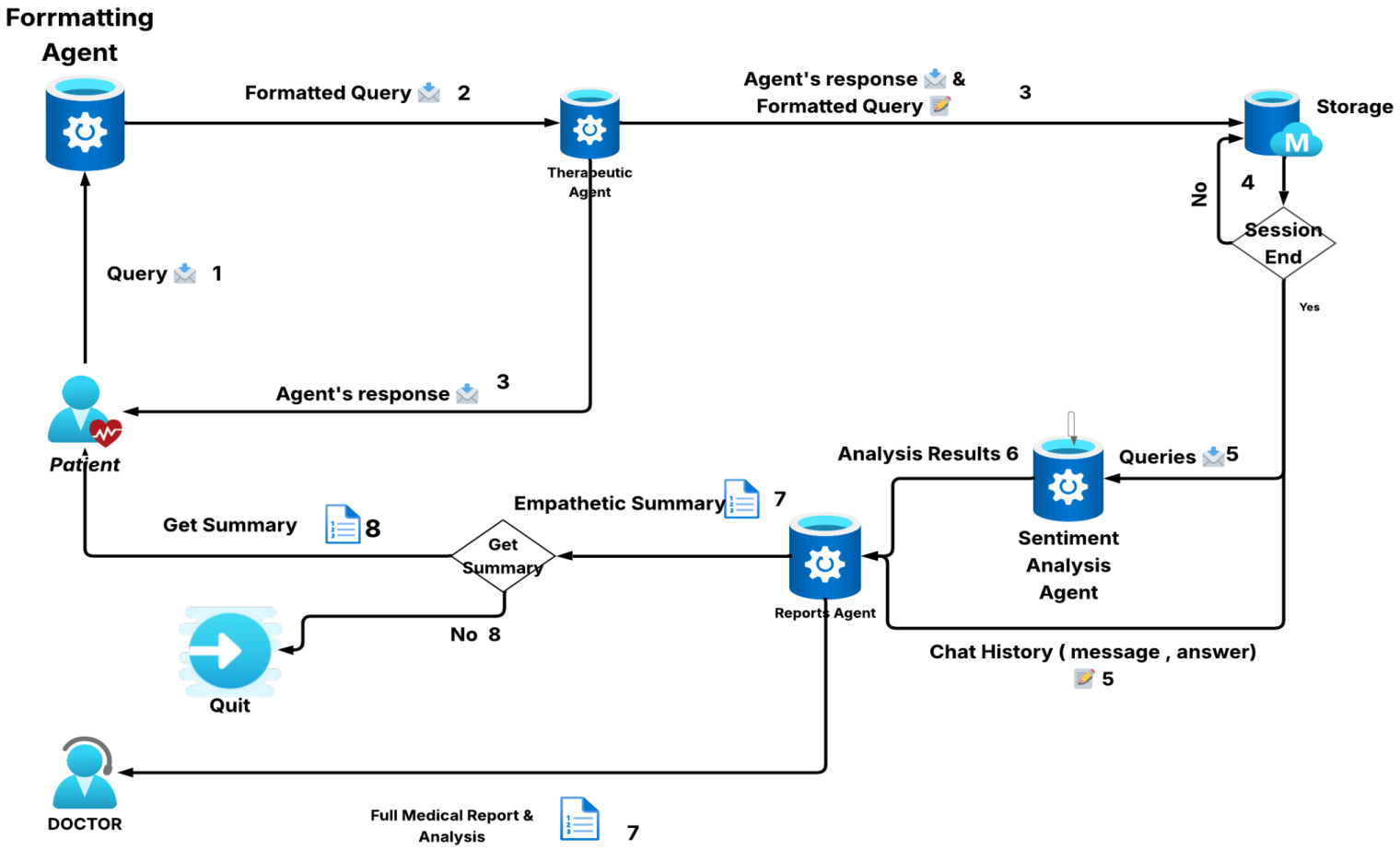


Figure 3.5 : System Workflow

This diagram presents the workflow of an AI-powered mental health agentic system composed of four key agents. The process begins when the patient submits a query, which is first cleaned and normalized by the **Formatting Agent** to ensure clarity. The refined input is then passed to the **Therapeutic Agent**, which generates empathetic and clinically informed responses using a fine-tuned LLM with Retrieval-Augmented Generation (RAG). These interactions are continuously stored and checked for session completion.

At session end, the entire chat history is forwarded to the **Sentiment Analysis Agent**, which assesses the emotional tone of the conversation. The results, along with the dialogue, are then processed by the **Report Agent**, producing either a personalized empathetic summary for the patient or a comprehensive medical report for clinicians. This structured workflow ensures accurate, emotionally aware support while offering valuable insights to healthcare professionals.

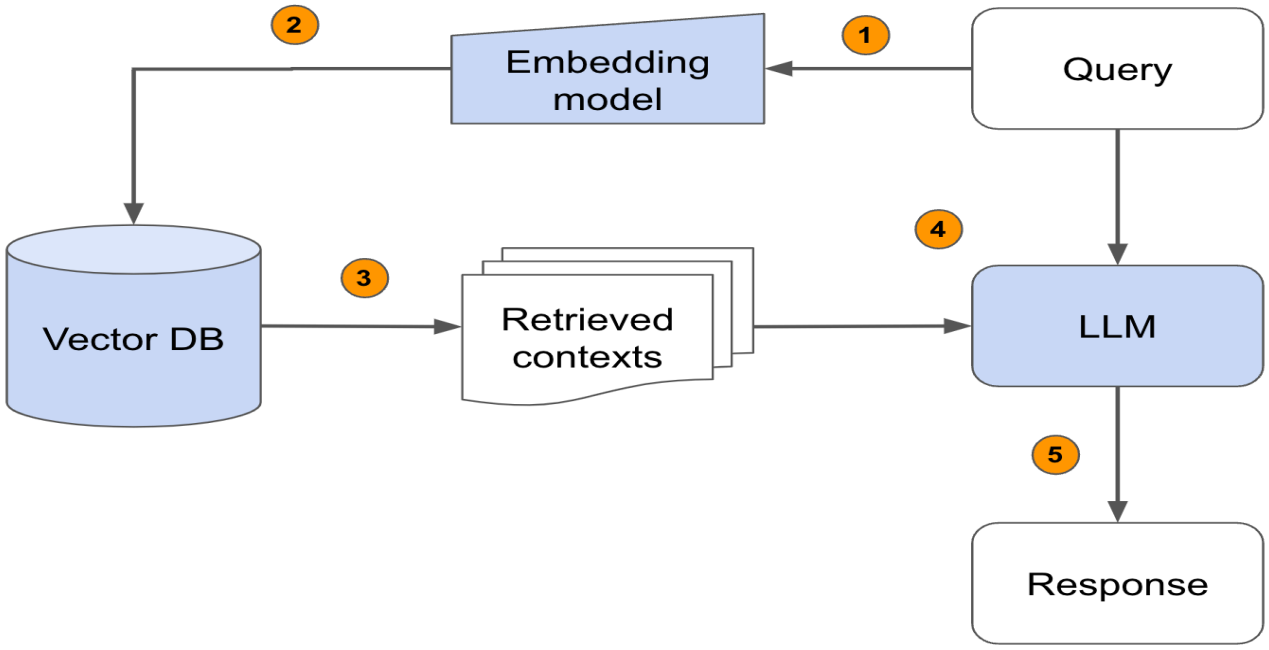


Figure 3.6: LLM + RAG Workflow

Retrieval-Augmented Generation (RAG) workflows showcase how AI can enhance decision-making by dynamically integrating real-time inputs with external knowledge. By retrieving relevant information from structured databases and synthesizing context-aware responses, these systems bridge the gap between raw data and actionable insights. This approach is particularly valuable in fields requiring precision and adaptability, as it combines the generative power of language models with verified, up-to-date sources. To ensure reliability, such systems benefit from built-in safeguards—like bias detection and human review mechanisms—that maintain accountability while leveraging AI's efficiency.

3.4. Implementations Specifics

3.4.1. Patient-Facing Mobile Application

The mobile application is specifically designed for patients, including parents, providing an intuitive and accessible interface for mental health support. Its core functionality revolves around direct conversational interaction with the Therapeutic Agent, enabling users to articulate their symptoms and concerns in a free-form manner. The system provides "empathetic summaries" of the user's health condition, accompanied by personalized recommendations and advice, which are distinctly differentiated from full medical diagnoses to prevent misinterpretation by non-medical professionals.

A key design feature is the integration of self-assessment quizzes, allowing patients to take up to

five quizzes per day. The results of these quizzes are processed by a dedicated agent that generates automated, personalized tips for mood regulation and distress management. These tips are then automatically displayed on the user's home page, offering proactive support and guidance. This inclusion of structured quizzes and personalized tips demonstrates a proactive, self-management oriented design philosophy, extending beyond reactive conversational support. It empowers patients with actionable insights for mood regulation, aligning with modern mental health approaches that emphasize patient agency and continuous, accessible support outside of formal therapy sessions.[78]-[81]

Beyond AI interaction, the mobile application facilitates direct communication through a chat functionality with the patient's assigned human doctor. This feature systematically provides the doctor with more context for their clinical work, enhancing the continuity of care. For parents, the application offers similar functionalities but with an adjusted architecture specifically designed to provide guidance and supportive methods for assisting their child's treatment journey.

3.4.2. Clinician-Centric Web App

The web application serves as a dedicated, comprehensive system for mental health professionals, designed to streamline clinical workflows and augment decision-making. Upon successful login, clinicians are presented with a centralized patient data overview, displaying key indicators such as patient status, names, genders, and appointment schedules.

The interface offers robust patient management capabilities, allowing clinicians to assign new patients, delete existing patient records, and review the complete history of past conversations with patients. A crucial feature is the ability to access patient mood data, as sent from the mobile application, and to review previous conversations between patients and the therapeutic agents. This provides critical contextual information, enabling a more informed clinical perspective. The platform also supports direct, secure chat functionality for real-time discussions with patients, enhancing communication efficiency. Furthermore, clinicians have the ability to manage appointments, including editing, deleting, and changing meeting dates.

Clinicians can directly interact with the LLM agents for general clinical knowledge and discussions, serving as a rapid consultation tool. A pivotal integration is the Report Agent, which empowers clinicians to generate detailed assessment diagnoses. This is achieved by synthesizing data from patient-doctor chat transcripts and the doctor’s personal notes. This functionality is explicitly designed to provide a better overview, enhance final reports, and

mitigate potential biases or assumptions in clinical judgment, thereby improving diagnostic accuracy and consistency. The doctor's system functions as a centralized clinical intelligence hub, integrating diverse data streams (patient moods, conversation history, personal notes) and LLM-driven analytical tools (Report Agent, general knowledge agent). This design facilitates a holistic, data-informed approach to patient care, potentially enhancing diagnostic precision and treatment planning by mitigating human cognitive biases.

Feature Category	Patient Mobile App Functionality	Doctor Web App Functionality	Key LLM Agent Involvement
User Authentication	Guest access, Authenticated access (Register/Login via Info/Google, select Client/Doctor).	Authenticated access (Login).	-
Core Interaction	Chat with Therapeutic Agent (text/voice).	Direct interaction with LLM Agents for general knowledge/discussions.	Prompt Formatting, Therapeutic Response, Sentiment Analysis
Patient Monitoring/Management	Take up to 5 daily quizzes; receive automated personalized tips for mood regulation.	Centralized patient dashboard (status, names, genders, appointments); Assign/delete patients; Review patient history/moods.	Sentiment Analysis (for mood data)
AI-Powered Support	Empathetic summaries of health conditions with recommendations/advice.	Generate detailed medical/therapeutic reports via Report Agent using patient-doctor chat/notes.	Therapeutic Response, Medical Report Generation
Reporting	Download empathetic summaries/rapport.	Download detailed medical/therapeutic reports.	Medical Report Generation
Communication	Direct chat with the assigned human Doctor.	Direct chat with patients; Meet a patient (Accept, Edit, Delete meetings).	-
Specialized Features	Parent-specific guidance and therapy aid for child's treatment.	Review patient conversations with Therapeutic Agents.	- therapeutic support

Table 3: Comparative Functionalities of Patient and Doctor Applications

3.5. Model Training and Assessment

3.5.1. Dataset Selection

The effectiveness and ethical integrity of LLM-powered mental health support systems rely heavily on the careful curation of datasets used during fine-tuning. These datasets are selected not only for their quality but also for their alignment with the sensitive and empathetic nature of mental health conversations.

We employed a diverse range of datasets to ensure our models exhibit supportive and compassionate behavior. These included:

- Friendly question answering
- Empathized responses
- Psycho-counseling conversations
- Psychological assessments and diagnostic data
- Symptoms and sentiment analysis

These varied datasets allow the model to learn nuanced therapeutic language and respond appropriately in emotionally sensitive contexts.

To enhance context-aware generation using Retrieval-Augmented Generation (RAG), we applied a strategic **data-splitting methodology.** A substantial portion of the training data was reserved exclusively for generating embeddings. This approach enables the system to retrieve more accurate contextual information during inference, thereby improving both the relevance of therapeutic responses and the clarity of clinical evaluations.

For fine-tuning, we adopted advanced domain-specific methodologies to tailor the LLMs to healthcare needs:

- Supervised fine-tuning: Trained the model on labeled datasets for tasks such as sentiment analysis and text classification, allowing precise control over output behavior.
- Transfer learning: Leveraged general-purpose pre-trained models and adapted them to mental health-specific tasks, enhancing performance with limited data.
- Parameter-efficient fine-tuning (e.g., LoRA): Adjusted only a subset of model parameters, significantly reducing computational demands while maintaining strong

performance. This approach is particularly effective for domains where high-quality data is limited.

- Continued pre-training: Extended the initial pre-training phase using domain-specific corpora before fine-tuning. This improved the model's comprehension of specialized psychological and clinical terminology.

By combining these fine-tuning strategies with carefully chosen, domain-relevant datasets, we were able to build empathetic and ethically aligned AI agents capable of supporting users in sensitive mental health scenarios. This specialization ensures our models go beyond generic language understanding, addressing concerns around bias, accuracy, and clinical reliability.

We utilized multiple datasets for fine-tuning and evaluation purposes, including those referenced in sources [70]–[75].

3.5.2. Evaluation Methods

To ensure the reliability and effectiveness of our models before deployment, we implemented a structured and multi-layered evaluation process. The dataset was carefully partitioned into subsets for training, fine-tuning, RAG embeddings, and a dedicated portion for testing. This setup helped maintain consistency and enabled robust, generalizable performance measurement across use cases.

Our evaluation strategy focused on two key components: semantic similarity and label-based prediction accuracy.

1. Semantic Evaluation

To assess the quality and empathy of generated responses, we employed multiple semantic similarity methods:

- LLM-based Semantic Evaluator: A separate language model was used to judge how well the generated responses aligned with ideal or reference answers. This provided a more nuanced, context-aware evaluation.

- Cosine Similarity (Sentence Transformers): We used Sentence Transformers to convert both the model output and reference responses into embeddings, then computed cosine similarity scores. This quantified semantic closeness in a vector space.

- WordNet Similarity (NLTK): Lexical similarity between generated and target words was

calculated using NLTK's WordNet interface. This was especially useful for detecting meaning overlap in psychologically relevant vocabulary.

These methods together ensured that our models produced responses that were not only technically accurate but also contextually and emotionally appropriate.

2. Label-Based Evaluation

For classification and diagnostic tasks (e.g., sentiment analysis or condition identification), we used standard supervised learning evaluation metrics:

- Accuracy
- Precision
- Recall
- F1-Score

These metrics were computed using predicted labels versus ground truth, helping us evaluate the models' ability to identify psychological patterns or categories correctly.

This evaluation methodology allowed us to validate not only the correctness but also the clinical appropriateness and empathic quality of the model's behavior—an essential requirement for applications in the mental health domain.

3.4 Conclusion

This chapter has presented the theoretical foundation of an AI-powered psychological support system that aims to enhance the delivery of mental health care. By strategically integrating specialized Large Language Models, the system is designed to improve the quality of therapeutic dialogue and increase the accuracy of psychological diagnostics, thereby addressing the growing demand for accessible, empathetic, and effective mental health support.

The system functions through two distinct but complementary modes that reflect the dual nature of mental health care. In the first, the Empathetic Support Mode, a fine-tuned Therapeutic Agent interacts with patients in a compassionate and emotionally intelligent manner. This agent, further enriched through Retrieval-Augmented Generation by incorporating psychological resources and previous patient data, is built to establish trust and create a safe environment that encourages patients to express themselves openly. In parallel, the Assisted Diagnosis Mode employs

LLM-generated insights combined with therapist observations to extract clinically relevant information and produce detailed diagnostic reports. This separation of roles enables the system to maintain ethical integrity while tailoring its behavior to meet both emotional and clinical objectives.

The system architecture supports seamless interaction between users, therapists, and AI agents, ensuring that conversation data and insights are shared, interpreted, and used constructively within therapeutic sessions. This deliberate design promotes both therapeutic empathy and diagnostic precision, ensuring that each component contributes meaningfully to patient care.

Ultimately, the system represents a significant step forward in the application of artificial intelligence to mental health. Its thoughtful combination of emotional support and diagnostic utility reflects a commitment to enhancing—rather than replacing—the role of human professionals in therapy. As such, it lays a solid foundation for the responsible integration of AI in psychological care, offering promising improvements in both patient experience and clinical outcomes.

This chapter is the basis of an actual AI-powered mental health tool that not only enables therapeutic discussion but also guides correct, informed psychological assessments, a significant achievement in AI applications for mental health.

Chapter 4:Operational Framework

4.1 Introduction

This chapter provides an overall description of a composition for an AI based medical assistant that is designed to deliver psychological support. Beginning with the tools and environments that were used during development, it details how the AI system is mainly made up of large language models (LLMs), Javascript and Typescript Frameworks ,Python and Google Colab provided us with the development environment and experimentation platform to train a Large Language Model with an artificial intelligence mindset. Libraries like LangChain handled the prompt management and logic of interaction, while Hugging Face and Kaggle provided pre-trained models and datasets to build the intelligence in the domain.

On the application side, a web-based dashboard was built using Next.js with Typescript and ShadCN UI kit components for psychologists to manage patient data and interact with the AI assistant. The web-based application was supplemented with a mobile application developed with bare React Native for the patient and their family for accessible, user-friendly interaction with the system. for The Database “supabase“ significantly simplifies the system's backend by providing a fully managed PostgreSQL database. It offers real-time data synchronization for instant updates between doctors and patients, and handles authentication securely out-of-the-box. This allows us to focus on building the core mental health support features rather than managing complex infrastructure.

The chapter, provided a description which includes integrating the components from data pre-processing to model orchestration and solving the front-end deployment for user engagement, and it ends with a performance evaluation on the utility, responsiveness, and reliability of the system in the application of herein MOD psychological health contexts.

4.2 Development Environment

4.2.1 Google Colaboratory (Colab)

Google Colab is a free, cloud-based platform that allows users to write and execute Python code within a Jupyter-notebook–style interface. It provides on-demand access to CPUs, GPUs, and TPUs, making it especially well‑suited for rapid experimentation, model training, and prototyping of large language models (LLMs). Colab’s seamless integration with Google Drive and its pre-installed Python libraries accelerate development workflows without requiring local hardware setup.

4.2.2 Python

Python is a high-level, interpreted programming language celebrated for its clear syntax, extensive standard library, and vibrant ecosystem. Its versatility spans web development, data science, machine learning, and scientific computing. In this project, Python serves as the primary language for model orchestration, data preprocessing, and backend services. Key Python packages (e.g., NumPy, pandas, and scikit-learn) facilitate data manipulation, while specialized libraries like FastAPI handle API endpoints. The Main Libraries the system is using are :

FastApi

A modern, high-performance web framework for building APIs with Python 3.7+ based on standard Python type hints. It's known for speed, automatic documentation (via Swagger), and ease of use.

Pyngrok

A Python wrapper for ngrok, which allows you to create secure tunnels to your local server. We used it to expose the colab's local FastAPI app to the internet (for testing & integrations).

Uvicorn

lightning-fast ASGI server for Python web applications. It runs your FastAPI app and handles asynchronous capabilities, making it ideal for handling high-concurrency requests.

4.2.3 LangChain

LangChain is an open-source framework designed to simplify the development of applications powered by large language models. It supports prompt management, tool chaining, memory handling, and integration with various LLM providers. In this project, LangChain was used to handle prompt orchestration and maintain conversational context between the user(doctor or patient) and the AI assistant. Its modular design allowed for scalable and customizable workflow definitions suited to psychological use cases.

4.2.4 Huggin Face & kaggle

The Hugging Face Transformers library and Datasets hub were utilized to integrate pre-trained models and access relevant data. Models such as GPT and BERT architectures were explored for fine-tuning and inference.

Kaggle was employed to retrieve annotated datasets specific to mental health, including text-based symptom descriptions, diagnostic labels, and therapy suggestions.

4.2.5 Typescript

TypeScript, a statically typed superset of JavaScript, was used across both the web and mobile codebases. Its type safety, autocompletion, and tooling integration significantly improved development reliability and maintainability. TypeScript enabled clear data models, reduced runtime errors, and enhanced collaboration across the development team by enforcing consistent interfaces between frontend and backend components.

4.2.6 Next.js and ShadCN UI

The web dashboard was developed using Next.js, a React-based framework optimized for server-side rendering and routing. It enables efficient data fetching and seamless user experience. All web components were written in TypeScript for improved developer experience and type safety.

ShadCN UI was employed to deliver a cohesive and accessible design system aligned with the needs of clinical workflows—such as navigating patient records, managing sessions, and reviewing chat histories.

4.2.7 React Native

The mobile application was built using React Native with TypeScript, ensuring shared logic and types between platforms. This enabled rapid development of a cross-platform solution (android and ios) for patients and caregivers. Features include AI chat, mood tracking, appointment reminders, and access to historical advice—all wrapped in a user-friendly mobile interface.

4.2.8 Supabase

Supabase is an open-source backend-as-a-service (BaaS) platform that provides instant APIs, authentication, and real-time database management. It was used to store and manage user data, chat histories, and session metadata. Built on PostgreSQL, Supabase offered robust data querying and role-based authentication, allowing secure access control between psychologists and patients. Its real-time sync features enabled seamless state updates in both the web and mobile applications, ensuring data consistency across platforms.

4.2.9 Git & Github

Version control and team collaboration were managed using **Git**, a distributed version control system, and **GitHub**, a cloud-based platform for repository hosting and project coordination. Git enabled developers to work in parallel across multiple features and environments, with branching strategies and pull requests ensuring code quality and integration consistency. GitHub facilitated team collaboration, issue tracking, release versioning, and continuous integration pipelines.

4.2.10 Vercel

Vercel is a cloud platform for frontend deployment and hosting, optimized for frameworks like Next.js. In this project, Vercel was used to deploy the web-based psychologist dashboard, enabling continuous deployment and rapid iteration. It offers serverless functions, automated builds, custom domain support, and global CDN delivery, ensuring fast, secure, and scalable performance for end-users.

Integrated with GitHub, Vercel automatically builds and deploys every commit pushed to the main or feature branches, enabling real-time preview environments and seamless collaboration between development and design teams. Its zero-configuration setup and tight integration with the Next.js framework made it an ideal choice for deploying our web application efficiently and reliably.

4.1.11 Lucidchart

Lucidchart was employed to design high-level system diagrams, user workflows, API data flows, and UI wireframes. As a cloud-based visual workspace, Lucidchart facilitated real-time collaboration, versioning of diagram changes, and integration with Google Drive and GitHub. These diagrams provided a shared visual language between developers, designers, and stakeholders, significantly enhancing communication, documentation quality, and development planning.

4.3 Implementation Details

4.3.1 Multi-Agent System Architecture

Agent	**Base Model Name**	**Description**	**task**	**Average Res Time**
Formatting Agent	vennify/t5-base-grammar-correction	Grammar LLm Handles the initial pre-processing step	Inputs Cleaning And Query Formatting	<0.2s
Therapeutic Agent	optimized LLaMA 3.2 Instruct model	Text-Generation LLm Takes the formatted user's query, converts it into a numerical vector ,searches & retrieves therapeutic content from the database to create an empathetic response	user Interaction And guide in treatment	<2s

Sentiment Analysis Agent	DistilBERT model	A lightweight model that processes all user inputs at the end of a session to extract a comprehensive sentiment profile.	Evaluates the emotional tone of the conversation which is then forwarded for reporting and documentation	<3s for each query
Report Agent	quantized Mistral 7B model	A Large And Powerful Quantized model That Takes Large organized data as input and engineered through prompts on how to deal with that data to generate the wanted results	For Professionals: Synthesizes a complete psychological report by incorporating user-therapist dialogue, sentiment analysis results, and optional clinician notes, to assist mental health professionals in their assessments. For Users (Non-Clinical): Generates concise summaries of user interactions, offering personalized advice, emotional validation, and recommendations.	<12sec

table4 :Names & Tasks of Each Agent

4.3.2 Evaluation

The evaluation process is critical in LLM's . It is necessary to check and supervise the evaluating first hand . This process was divided into two parts because we used multiple LLM's with multiple strategies. in which we decided to go for :

4.3.2.1 Fine-Tuned Models Evaluation

The fine-tuned model were evaluated by comparing both a base-model which was not fine-tuned to check if the fine-tuned model offers a more empathic question answering behaviour which was the case.

and another method was comparing it directly to the dataset answers To ensure that the text-generated answers were relatively sympathetic and chatty. allowing the user to trust and be more open to the model.

these next the config were used to generate the answers aiming for a good balance between creativity and performance

```
generation_config = GenerationConfig(
 max_new_tokens=500,      # Limit on the length of the output
 temperature=0.7,         # Controls randomness/creativity
 do_sample=True,          # Enables sampling over greedy decoding
 top_p=0.9,               # Nucleus sampling — limits to most likely tokens
 repetition_penalty=1.1,                  # Penalizes repeated words
)
```

questions array = [

```
"What if I'm just broken and can't be fixed?",
 "Sometimes I have thoughts that scare me.",
 "What's the point of any of this?",
 "I keep thinking about dying even though I don't want to die.",
 "I feel like I'm losing control over my own mind.",
 "Is it okay that I want to take a break from everyone?",
```

]

aspect	Response 1(Fine-Tuned Llama)	Response 2 (Base Model Without fine-tuning)
on	I hear you. It sounds like there's a deep-seated sense of hurt	I understand you might feel that way, but you don't need to.
Exploration	Can you tell me more about what's going on for you?	What do you mean by this concept?" (Overly analytical)
Collaboration	We can work together to find ways to heal.	You should consider seeking treatment." (Directive)
Non-Judgmental Tone	Your feelings are valid.	It's normal to feel that way when you struggle." (Labels)
Hope & agency	You don't have to be defined by this.	Addiction is hard to overcome on your own." (Assumptive)

Table5 : Responses to the same queries from both models.

The Code and all tools used for Evaluating is provided in ***this link here***

given the table above we can notice that the fine-tuned model responses are significantly more empathetic in comparison to the base-model. the main differences that puts our fine-tuned model as the better are :

- **It prioritizes emotional validation** before problem-solving.
- **It invites users to share** without judgment.
- **It uses a collaborative tone** ("we," "together").
- **Affirms agency** while acknowledging pain.
- **Emotional Safety**: Creates a space where users feel heard, not analyzed.
- **Person-Centered**: Focuses on the user's unique experience (e.g., *"What makes you think, 'I am broken'?"* vs. *"It's not about you."*).
- **Hope-Oriented**: Balances validation with actionable support (*"We can work on coping strategies"*).

while the base-model is acceptable but does sound more **transactional (less warmth) ,** also its **Focuses on fixing** rather than listening. The answers have a **stigmatizing language** in some responses.as well as in some cases it goes as far as **labeling the users** ("mental health issues") prematurely.

4.3.2.2 RAG evaluation

This evaluation focuses on the performance and functionality of the Retrieval-Augmented Generation (RAG) system designed to answer psychotherapy-related questions. The system was implemented in Google Colab using tools from Hugging Face and FAISS, with the goal of combining semantic retrieval with language generation for producing contextually relevant and accurate answers.[76]

1. Dataset and Setup Evaluation

The system operates through two multiple embedding files. The main dataset for this project, PsychoCounsleling.npy , contains a large number of context-response pairs representing actual therapy content. Each item includes a question (the context) and its corresponding answer, which allows the system to create a base set of knowledge to run the retrieval against.

For the evaluation, the system relied on a separate CSV file, Evaluation_Sheet.csv, which contains test questions with their true answers. The system loops through each question in the

CSV and runs the retrieval method on the Psychology.json file using FAISS vector similarity search to retrieve the top-k (e.g., 5) items for each question.

The items retrieved are then passed with the evaluation question to the language model, which generates an answer to the question. The generated answer was then compared to the expected answer in the CSV using cosine similarity to produce two embedding sentences to quantify semantically how close the generated answer was to the ground truth answer.

2. Retrieval Functionality

The retrieval step processes context dataset data and incoming queries using the paraphrase-MiniLM-L6-v2 model from SentenceTransformers to make dense vector embeddings. Then, fast and efficient similarity searches take place with FAISS.

This method worked well for retrieving the therapy response that had been generated, especially when the user question closely matched the wording in the context dataset. However, in cases when the questions were ambiguous or reworded, there were times the retrieval returned less relevant data, negatively influencing the quality of the generation.

3. EGeneration Phase Analysis

The generation process used the microsoft/phi-2 model, which was accessed via the Hugging Face Hub. The microsoft/phi-2 model is a causal language model designed to produce coherent and contextually appropriate language as it is fine-directed for generation based on the retrieved therapy notes and the evaluation question.

After the top-k relevant context entries are retrieved, they are passed as well as the evaluation question to microsoft/phi-2, which generates the answer based on the provided context in one attempt to provide a knowledge-informed, empathetic, and contextually accurate response.

The quality of the answers depend on the relevance of the top-k retrieved entries. The weaker or more off-topic the retrieved information, the more generic or off-based the generated answer tends to be.

4.Output and Result Storage

Every generated answer along with its cosine similarity compared to the correct answer is saved into Evaluation_Sheet.csv. This creates a visible and organized analysis of how your system performed across all questions that you evaluated.

5.Conclusion

The evaluation of the RAG system, which performed semantic retrieval plus the microsoft/phi-2 model for answer generation, resulted in a cosine similarity score of 0.700125716 which reflects a moderate fit of the responses and answers created. The system performs above standard, though there is room for improvement perhaps through improving the retrieval accuracy and model tuning that will provide even better quality and consistency of the generated answers.

As For the Diagnoses Report The LLama 3.2 was provided with copies of mental health reports to add to the certainty of having an accurate formatting to all the Notes and observations extracted as well as having the DistelBert which is used for mental health diagnoses and emotion Detection for more precision.

4.3.2.3 Overall Evaluation

We Compared the The generated Responses Similarity with the true DataSet Answers that's provided from psychotherapists and professionals for each model .we also choose other models for more general comparison.

The Average Similarities Were Calculated Using the **Semantic Similarity Evaluation**

In this evaluation approach, we measure how semantically similar the AI-generated response is to a reference (ground-truth) response. Instead of relying on exact string matching (like BLEU or ROUGE), we use **dense vector embeddings** to assess meaning-level similarity .

the model's response and the reference answer are:

- Encoded into 384-dimensional embeddings using the embedding model.

Compared using **3 Methods of similarity, cosine similarity , python nltk and wordnet Semantic similarity , and semantic similarity calculated by an LLM.**

Models	Llama 3.2 instruct base Model	Jarvis (Llama 3.2 instruct + Fine-tuning)	microsoft/phi -2 + RAG	Jarvis + RAG	GPT2 -LARGE
cosine similarity	0.635798	0.7635196	0.700125	0.81137	0.4863687
LLM Similarity	0.65983	0.78567	0.72123	0.83456	0.52345
WordNet nltk Sim	0.58765	0.71234	0.69876	0.78432	0.51678

Table6 : Comparison between multiple LLMs on the same queries

These were The results gotten from a 1000 rows Sample, the excel sheets are included in the following : [77]

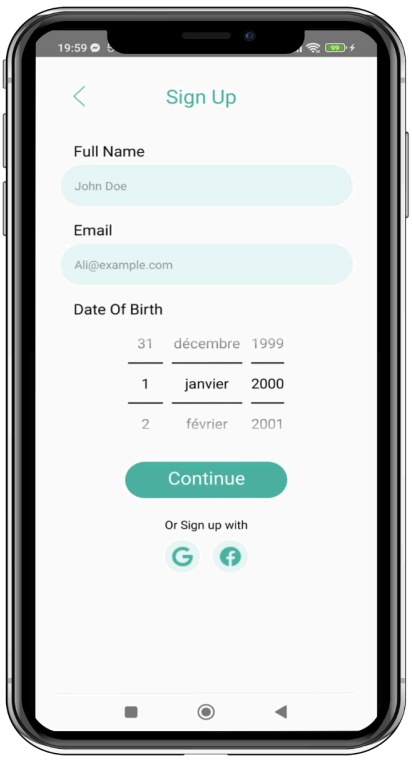
4.4 Graphical User interface

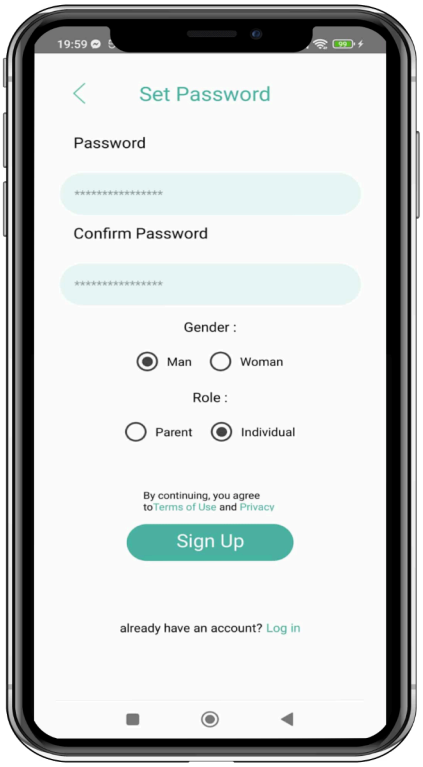
The design of the graphical user interface (GUI) is critical; it allows users (doctors, patients, parents, etc.) to view and interact with the system easily and seamlessly. In the final implementation, we had two distinct interfaces: a responsive web application for the psychological health providers, and a mobile app for the patients and their caregivers. Each interface was designed specifically with the user in mind to make the technology clear, easy to use, and easily engaging.

4.4.1 Mobile Application

4.4.1.1 Signup :

The mobile application begins with a streamlined and user-friendly signup process designed to collect essential user information while ensuring accessibility across different user roles. New users are prompted to provide their full name, email address, date of birth, gender, and define their role as either a **Patient** or a **Parent**. 4.1 SignUp screen

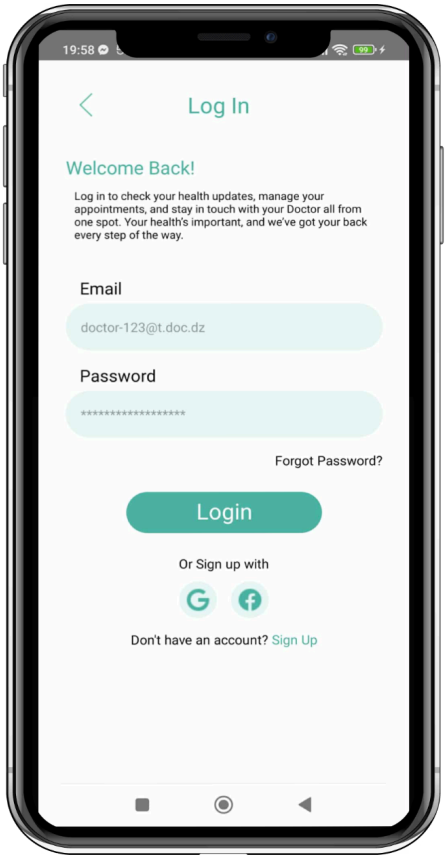




To ensure account security, users must also enter and confirm a secure password. Once the form is completed, all data is validated locally and then securely transmitted to the Supabase backend, where it is stored in the appropriate user table. This structured registration flow lays the foundation for a personalized in-app experience based on the user's profile and role.

4.4.1.2 Login:

To access the mobile application, the user begins with the Login process. This step requires entering a valid email and password to authenticate the account. If the user has forgotten their password, a "Forgot Password" option is available to initiate a reset through email verification. For users who prefer a faster or alternative method, the app also provides the option to sign in with Google or Facebook. This ensures flexibility and ease of access, especially for those who haven't registered through the standard sign-up process.

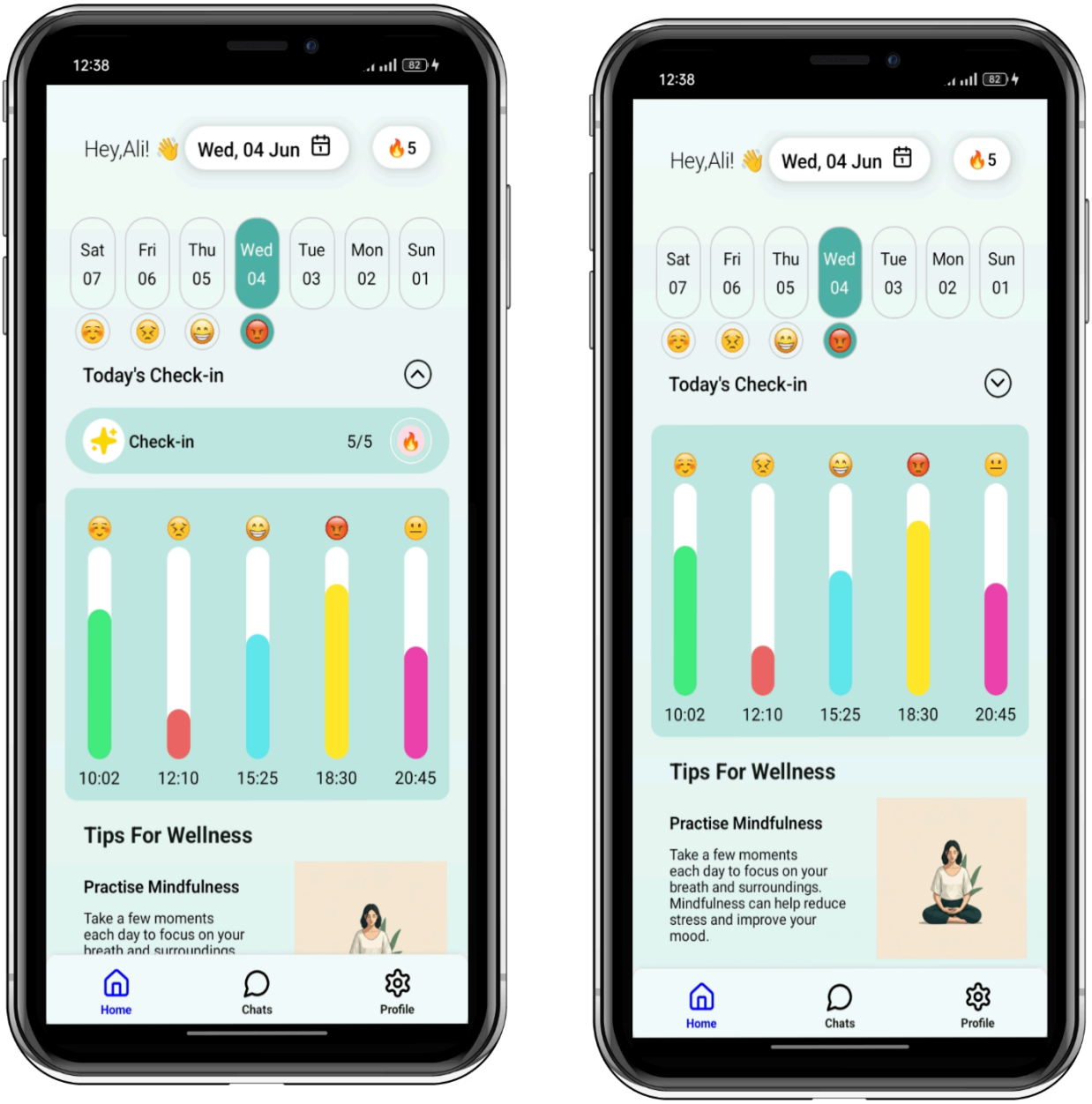


4.2 Login screen

4.4.1.3 Home Page:

Upon logging into the application, users are welcomed on the Home Page, which serves as the main hub for daily engagement and emotional tracking. At the top, the app greets the user by name, displays the current date, and shows a motivational streak counter represented with a flame icon to encourage consistent usage. Just below, a horizontal mood tracker allows users to navigate through the week and view their emotional states across different days using expressive

emoji icons. The “Today’s Check-in” section provides a timeline of mood entries for the day, visually represented with colored bars and emojis that reflect the emotional state at specific times. This enables users to monitor their emotional fluctuations throughout the day. The home screen also offers Wellness Tips, such as practicing mindfulness, to guide users toward healthier mental habits. Overall, the Home Page is thoughtfully designed to promote daily emotional awareness and personal growth in a clear and engaging manner.



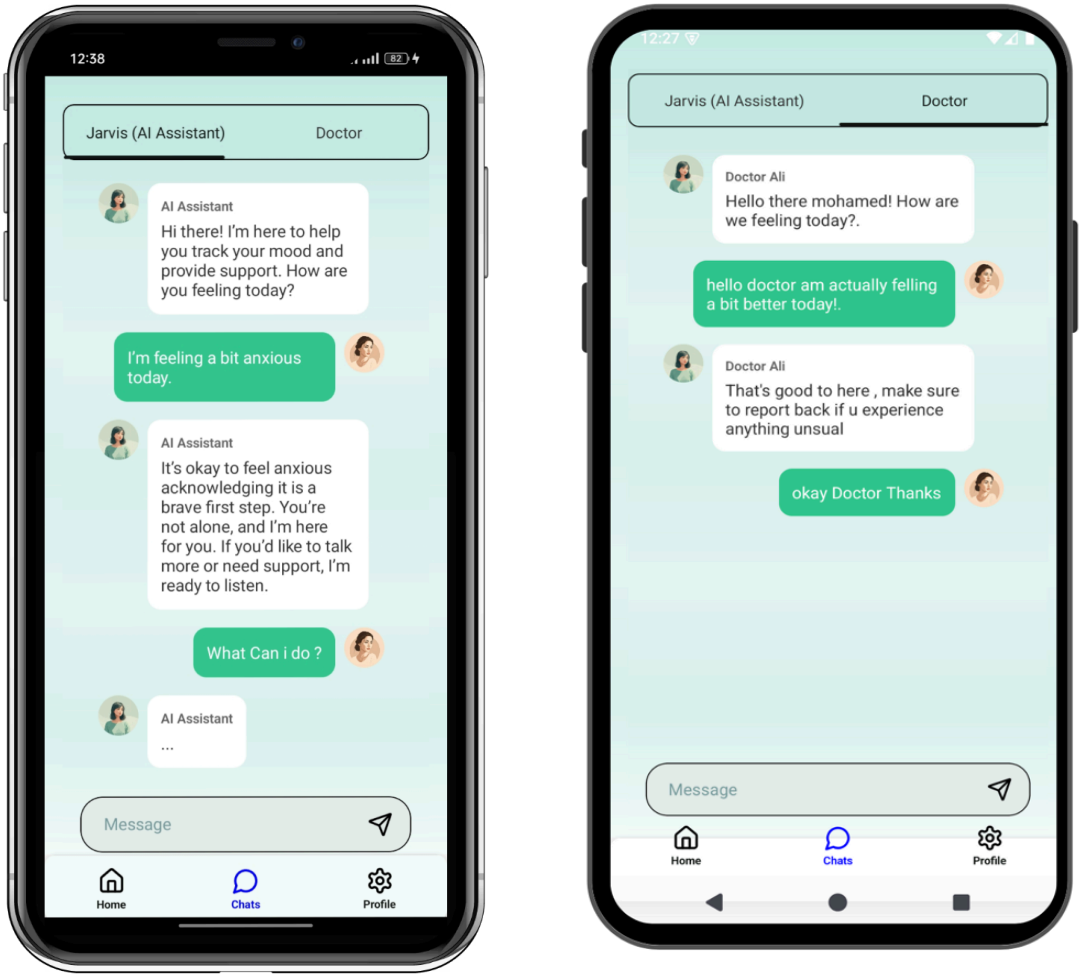
4.3 Home screen

4.4.1.4 Chats :

The app features two chat interfaces: one with Jarvis, an AI assistant for emotional support, and

another with the user's doctor for medical follow-up. Users can share their feelings with Jarvis, who responds empathetically to help them cope, while also maintaining communication with a real doctor when needed.

To ensure high-quality interaction, all messages are pre-processed—including grammar correction and user-type identification (e.g., parent, patient)—before being sent to the AI assistant. This ensures clear, personalized, and helpful responses.



4.4 Ai & Doctor Chat screen

4.4.1.5 Settings :

The Settings screen allows users to customize their app experience and manage key preferences. It includes the following options:

Notifications: Toggle push notifications on or off to stay informed about updates or remain undisturbed.

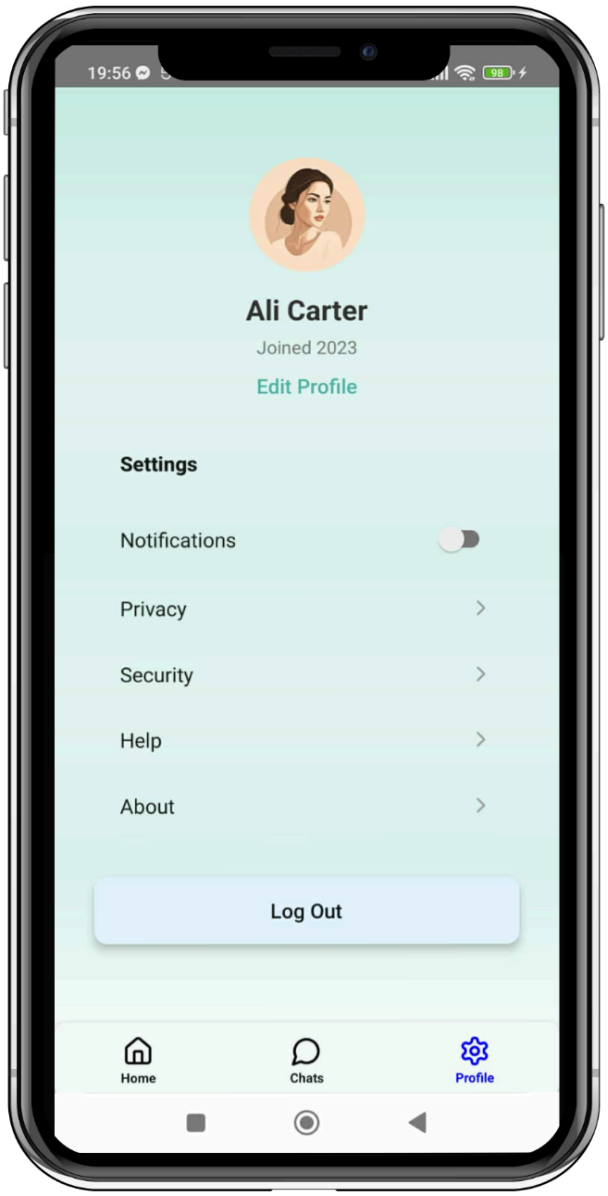
Privacy: Manage data sharing preferences and view the app's privacy policy to understand how user information is handled.

Security: Update your password and configure security settings to keep your account protected.

Help: Access support resources and FAQs for quick assistance with using the app.

About: Learn more about the app, its purpose, and version information.

logout: Securely sign out of your account from the device.



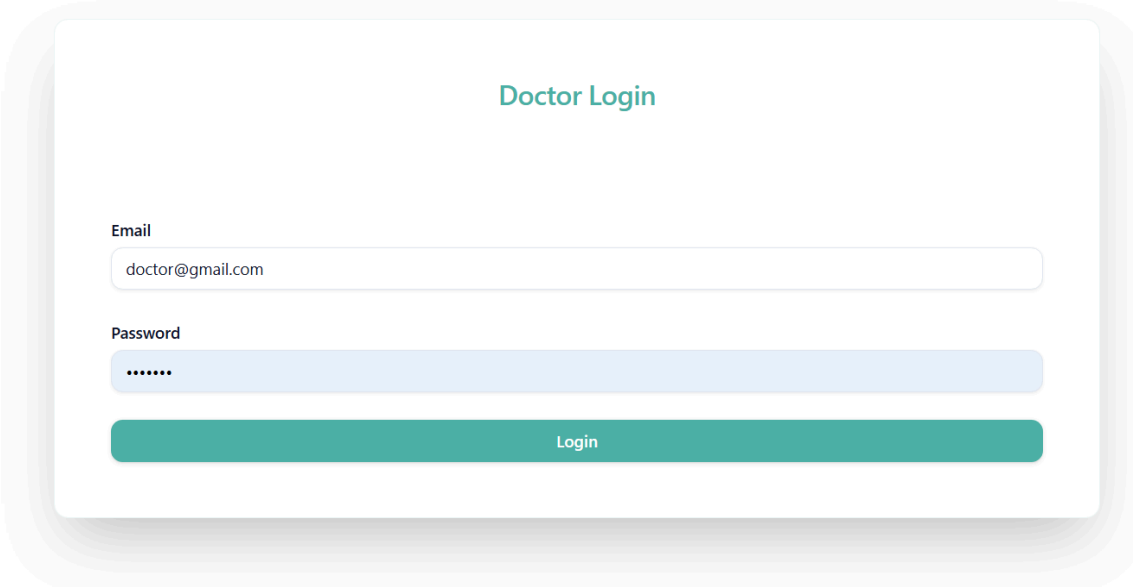
4.5 Profile screen

4.4.2 Web App

The Psychology Dashboard is a modern, fully responsive web application designed to support mental health professionals and their patients. It features an intuitive dark and light mode toggle, ensuring comfortable use in any lighting condition. A fixed sidebar allows seamless navigation between key sections of the platform, including appointments, patient records, AI chat, and settings. Each page includes a top navigation bar displaying the section title, user avatar, and a role indicator, along with a dedicated search field for quick access to relevant content. This streamlined layout enhances usability, making daily clinical and administrative tasks more efficient and user-friendly.

4.4.2.1 Login :

The Doctor Login page is a secure authentication interface developed using Next.js and integrated with Supabase for backend authentication. It allows registered doctors to log in using their email and password. The form validates user input and submits credentials via Supabase's Auth API. Upon successful login, an access token (JWT) is issued and used to access protected routes. All data is transmitted over HTTPS, and the token is included in the Authorization header (Bearer <token>) for subsequent requests. This ensures only authorized users can access sensitive areas of the system.



4.6 Login dashboard

4.4.2.2 Home Page (Overview):

The **Home** or **Overview** page provides a comprehensive dashboard that summarizes key statistics and user data in a visually interactive format. It is fully responsive, ensuring optimal display across devices of all sizes.

1. Diagnostics Statistics

This section visualizes the psychological diagnostic categories of patients over the last six

months. A donut chart is used to display the distribution, helping identify the most common mental health issues such as stress, anxiety, depression, and more.

A growth percentage indicator (e.g., +5.2%) reflects the change in patient diagnostics for the current month, indicating trends in engagement and mental health awareness.

2. Genders Statistics

This section provides insights into the gender distribution of patients for the current month. It includes a similar donut chart representation, making it easy to compare groups visually. In this example, 295 patients are shown with a +8.7% weekly growth.

3. Monthly Patients & Sessions

A bar chart illustrates patient activity and session count across the months from January to June. Each bar compares two datasets—patients and sessions—enabling clear tracking of app engagement and service utilization. A green trend indicator below the chart highlights percentage change for the current month (e.g., “Trending up by 5.2%”).

4. Geographic Distribution (Map of Algeria)

A color-coded map of Algeria highlights patient distribution by region. Each province is represented with a distinct color, and the patient count is displayed next to the region's name. For example: Béjaïa: 400 patients , Médéa: 300 patients , Adrar: 200 patients….



4.7 Overview Page

4.4.2.3 Appointments :

The Appointments screen provides patients and professionals with a well-organized monthly calendar to track all scheduled therapy sessions, consultations, and follow-ups. This feature improves clarity, planning, and time management for mental health providers.

1. Monthly Calendar View

A full-month calendar layout displays all appointments scheduled for the current month (e.g., May 2025). Each appointment appears as a colored tag indicating the session type or mental health topic, such as:

Therapy Sessions , Cognitive Behavioral Therapy (CBT) , Depression Therapy ,Family Counseling , Medication Follow-ups , Emergency Sessions (e.g., Panic Attack)

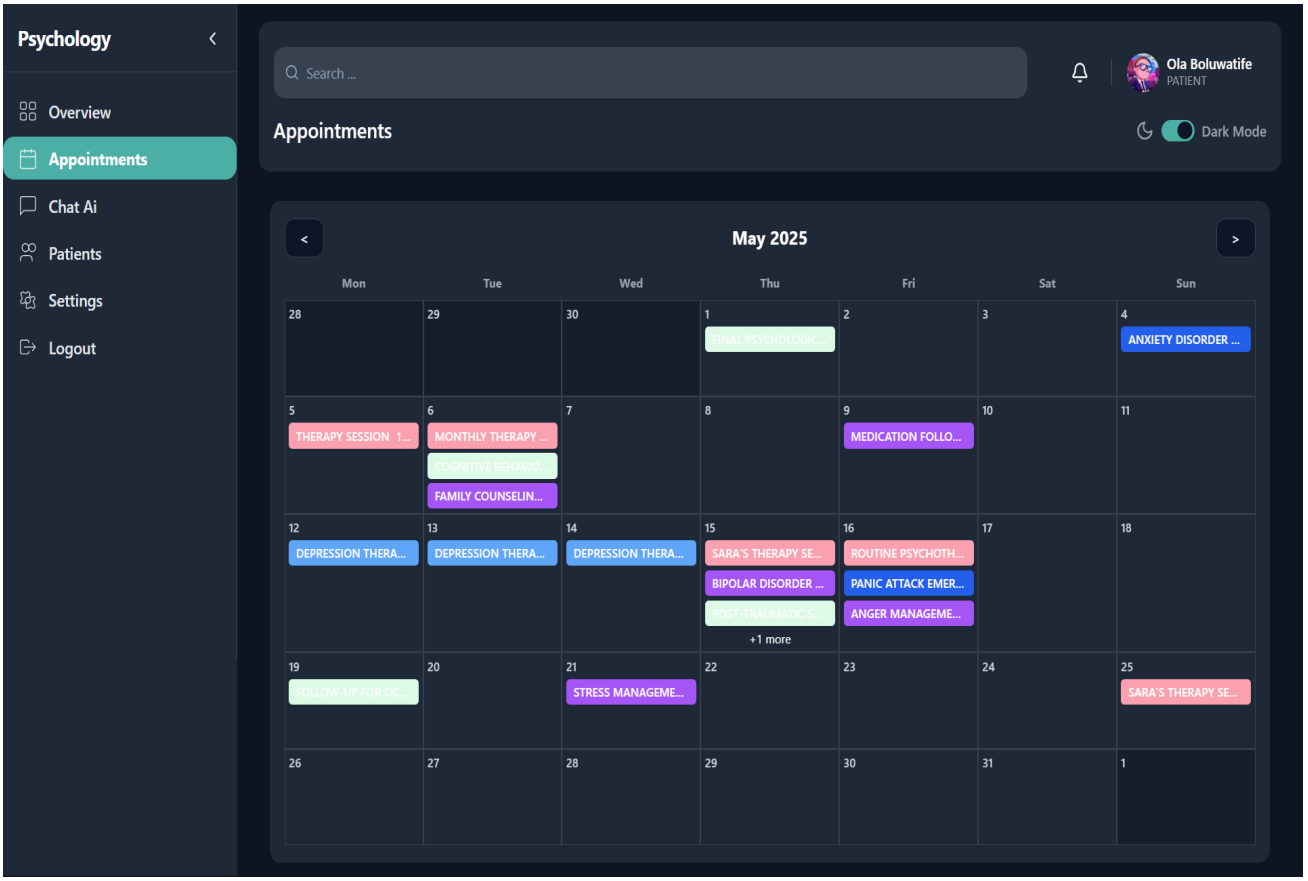
This layout helps users quickly understand their schedule, with color-coded tags for easy

identification of session types.

2. Session Diversity

A wide variety of psychological topics are reflected in the calendar, showing the platform's ability to accommodate different mental health needs. Common session types include:

- Depression, Anxiety, OCD, and Bipolar Disorder sessions
- Routine psychotherapy and follow-ups
- Specialized sessions like anger management and post-traumatic stress therapy
- Named sessions such as "Sara's Therapy Session" for personalized or recurring care



4.8 Appointments Page

4.4.2.4 Chat Ai :

The Chat AI feature integrates an advanced language model to support psychologists in real time. It acts as a virtual assistant to enhance clinical decision-making and administrative efficiency.

1. AI-Assisted Diagnostics

The Chat AI provides intelligent support to doctors by analyzing patient symptoms, behavior descriptions, and historical records. It assists in:

- Suggesting possible diagnoses based on patient input
- Identifying signs of conditions like depression, anxiety, …and more
- Recommending psychological assessments or next steps
- Reducing cognitive load by narrowing down differential diagnoses

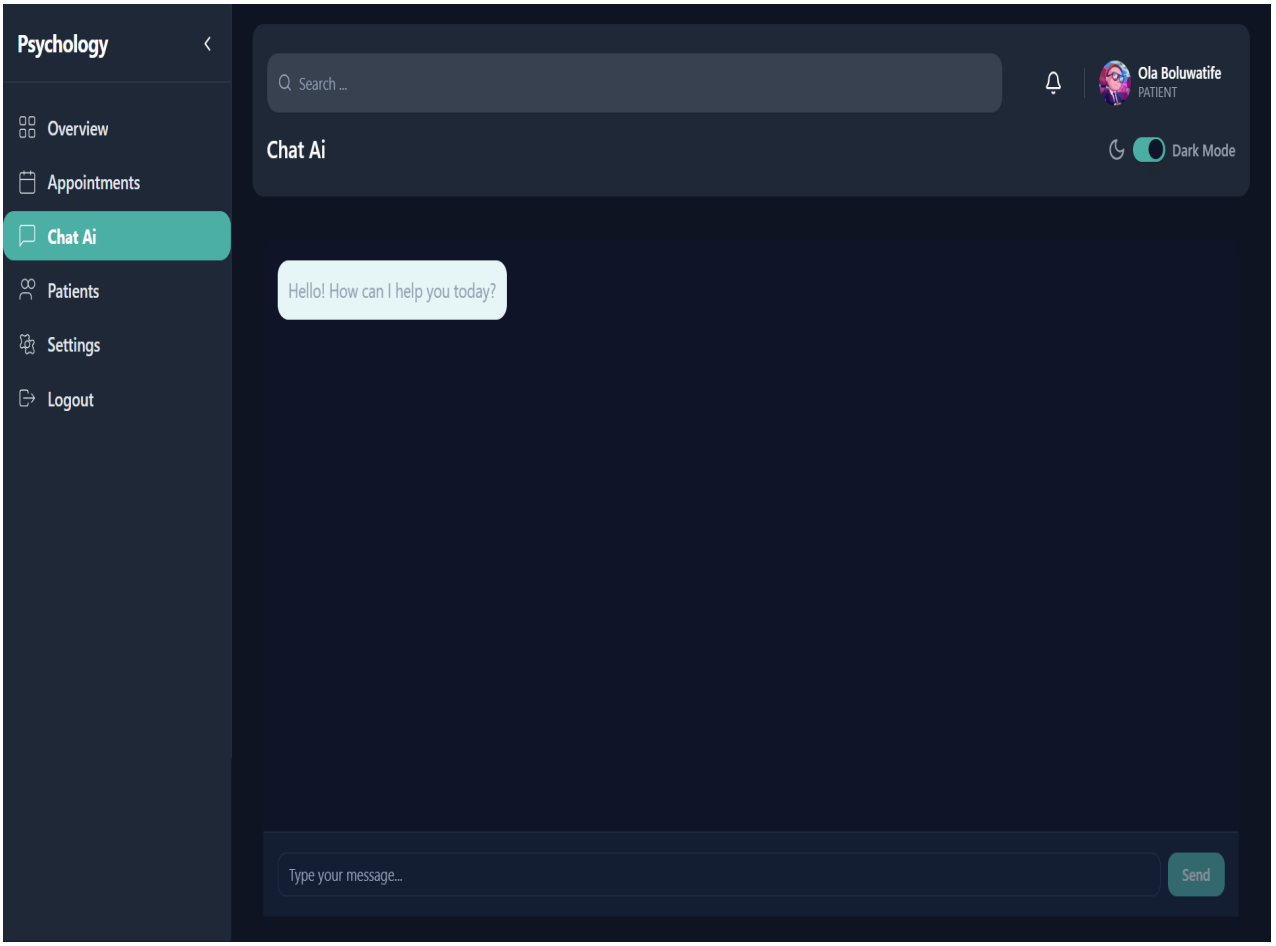
2. Automated Patient Reports

The system can generate comprehensive psychological reports summarizing:

- Patient history and session notes
- Detected behavioral patterns or emotional triggers
- Progress across sessions
- Suggested therapeutic approaches

3. 24/7 Virtual Mental Health Assistant

Available anytime within the app, Chat AI responds to clinician questions, provides therapeutic suggestions, and explains psychological concepts — making it a powerful educational and operational tool for mental health professionals.

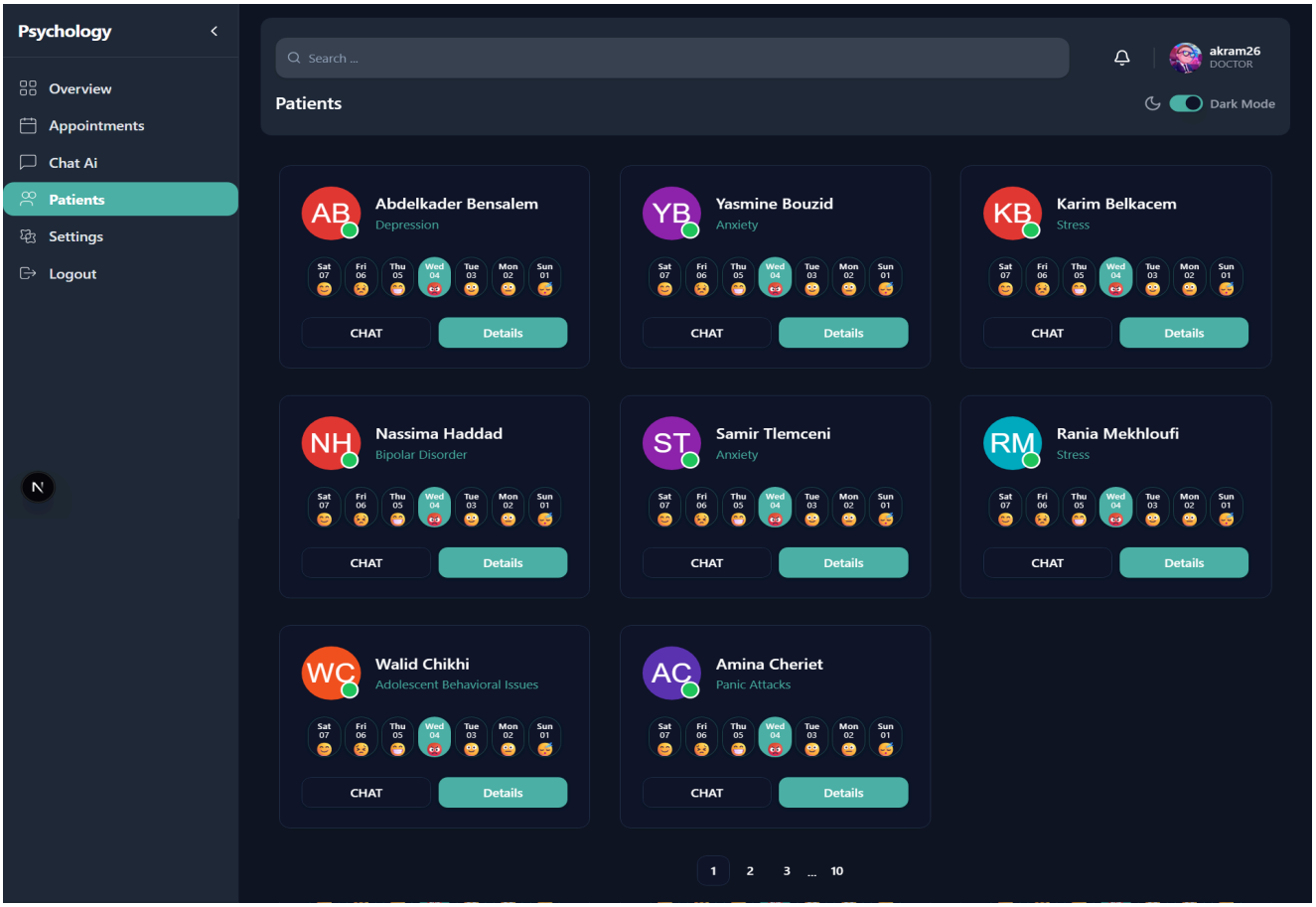


4.9 CHAT -AI Page

4.4.2.5 Patients :

The Patients module serves as the central hub for managing psychological patient profiles in a structured and intuitive layout. Each card represents an individual undergoing therapy or psychological treatment and offers direct interaction features for clinicians.

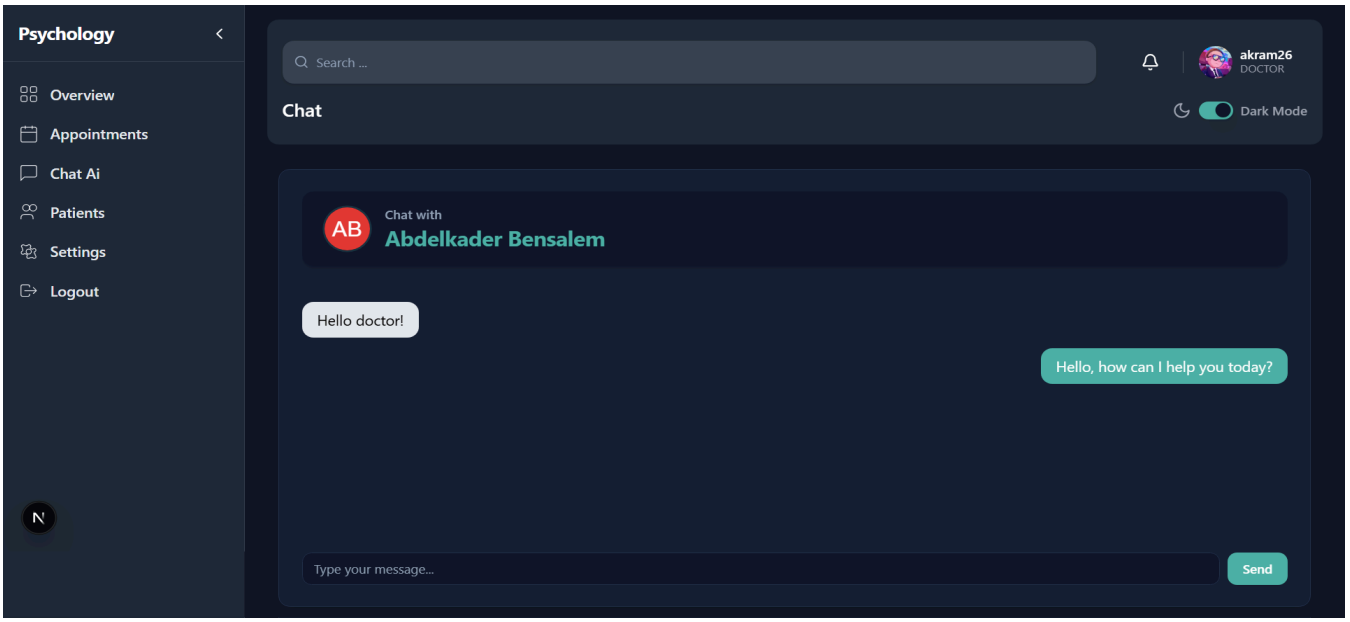
Each patient card also visualizes a weekly emotional trend using emojis corresponding to daily mood states. For instance, a sequence may include: 😊 (happy), 😣 (anxious), 😁 (joyful), 😡 (angry), 🙂 (content), 😐 (neutral), and 😴 (fatigued). This feature provides clinicians with a fast and intuitive overview of the patient's emotional fluctuations across recent sessions.



4.10 Patient Page

1. Direct Patient Communication (CHAT Button)

The Chat button opens a real-time messaging interface between the assigned psychologist and the patient.

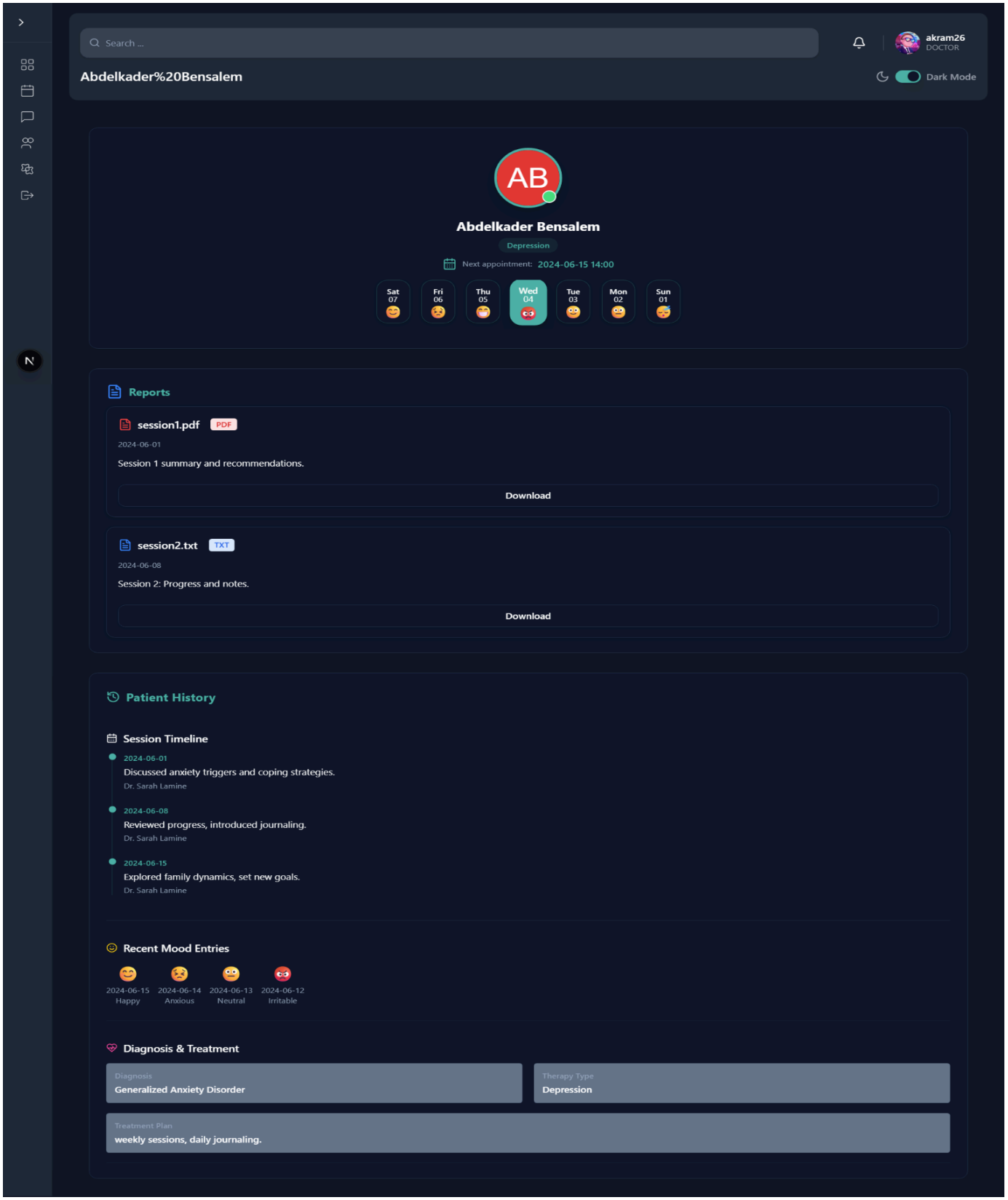


4.11 Patient Chat page

2. **Comprehensive Patient Summary Access (Details Button)**

Clicking the Details button opens the patient's comprehensive psychological profile, offering clinicians a well-organized and data-driven interface to support ongoing care.

- Patient Overview: The top section displays the patient's full name, current psychological diagnosis, and the date of their next scheduled session—ensuring immediate access to essential clinical data.
- Emotional Mood Tracker: A visual tracker summarizes the patient's emotional states over the past seven days using emojis, providing a quick, intuitive overview of recent mood fluctuations.
- AI-Generated Reports: This section contains intelligent summaries automatically created following chat interactions between the patient and the system. These reports are sent directly to the assigned clinician and include behavioral insights, psychological observations, and clinical recommendations.
- Session History: A chronological timeline logs all past therapy sessions, highlighting key discussion points, therapeutic strategies used, and evolving goals within the treatment plan.
- Mood Entries Log: Recent mood entries are time stamped and labeled with emotional indicators such as 😊 Happy, 😣 Anxious, 😐 Neutral, or 😡 Irritable, offering context to behavioral changes and progress.
- Diagnosis & Treatment Plan: This final section outlines the current diagnosis, the type of therapy being followed, and any active interventions—such as weekly therapy sessions or prescribed journaling exercises.



4.12 Patient Details page

4.4.2.6 Settings :

The Settings page allows medical professionals (psychologists) to manage their personal

information and application preferences.

1. Account Details Section

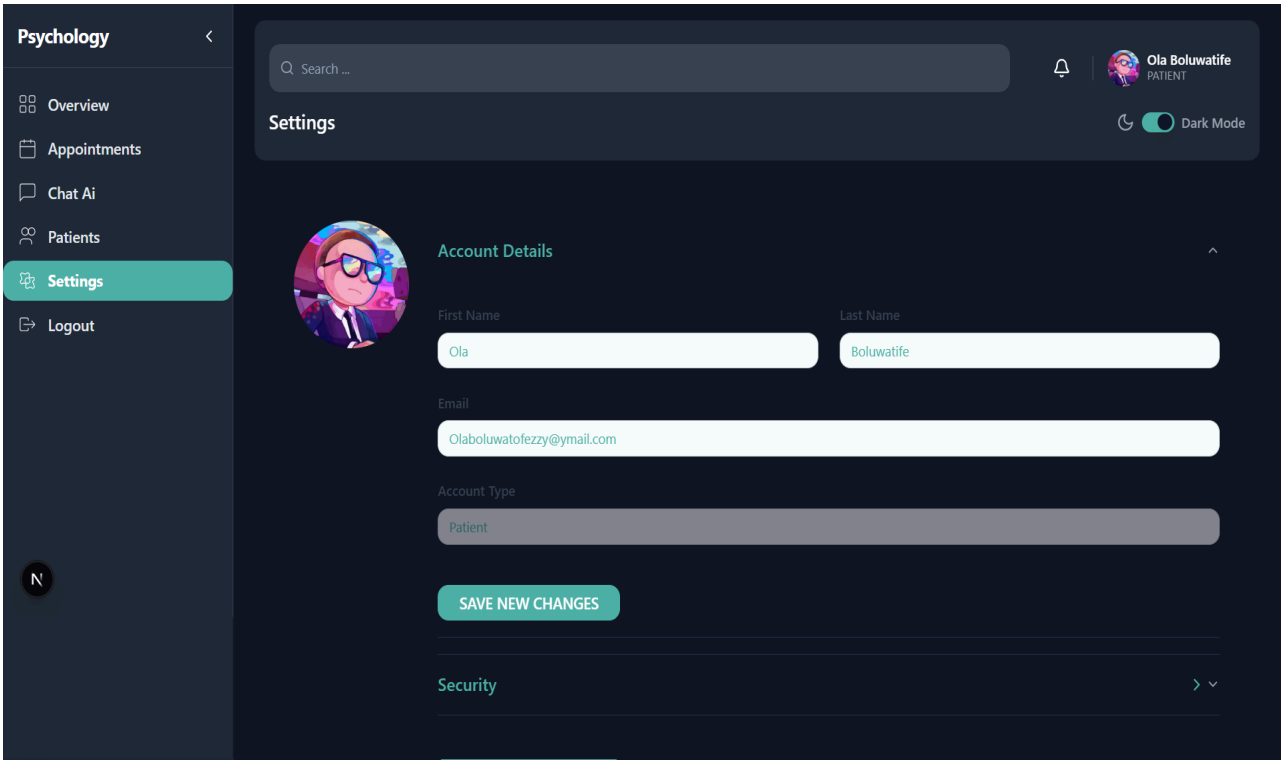
- Users can view and update the following information:
- First Name / Last Name: Editable fields for personalization.
- Email Address: Used for notifications, recovery, and account login.
- Account Type: Displays the user role (e.g., Patient, Doctor, Admin). This field is non-editable for security and role-based access control.

2. User Avatar

A customizable profile image is displayed to enhance user identity and personalization.

3. Security section :

- password change
- two-facture authentification (2FA)
- recent login activity



4.13 Doctor Profile page

4.5 Conclusion

This chapter presented the comprehensive operational framework of our AI-based psychological assistant system. By integrating powerful tools, modern programming languages, and cloud-based platforms, the system was built to be both technically sound and clinically effective. The development environment—featuring Python, Google Colab, LangChain, and Hugging

Face—enabled efficient model training, evaluation, and orchestration of various AI agents. On the application side, robust and responsive interfaces were created using TypeScript, Next.js, and React Native, ensuring usability across both web and mobile platforms.

Back-end infrastructure, supported by Supabase, allowed for real-time synchronization, secure authentication, and scalable data management—all critical for delivering safe and seamless interactions between patients and mental health professionals. Moreover, the implementation of a multi-agent LLM system allowed for specialized task handling, from sentiment analysis to empathetic conversation generation and professional report synthesis.

Thorough evaluations, including both fine-tuned model assessments and RAG-based retrieval analysis, demonstrated the system's capacity to produce empathetic, context-aware responses with high semantic fidelity. By comparing multiple models and embedding strategies, we validated the efficacy of our fine-tuning approach and the role of semantic search in improving the relevance of generated responses.

Finally, the dual-interface design—tailored to meet the needs of both clinicians and patients—ensures a holistic digital mental health ecosystem. It emphasizes emotional safety, accessibility, and clinical utility, making the platform not only a technological achievement but also a meaningful step toward advancing psychological care through AI.

This foundation sets the stage for future scalability, improved personalization, and enhanced diagnostic support, all of which will be explored in the system evaluation and future work discussions in subsequent chapters.

General conclusion :

In conclusion, this project presents a comprehensive and modular AI-based system designed to support psychological well-being by integrating multiple language models and modern development tools into a cohesive architecture. The system is built around the concept of multi-agent collaboration, where each agent plays a distinct role — from natural language formatting and therapeutic interaction to emotional sentiment analysis and clinical report generation. This design allows the system to respond not only with technical accuracy, but also with contextual sensitivity, making it suitable for both end-user support and professional mental health applications.

By leveraging advancements in large language models, prompt orchestration (LangChain), and scalable infrastructure (Supabase, Vercel, etc.), the platform offers real-time, accessible, and structured mental health assistance through both web and mobile interfaces. The implementation demonstrates how artificial intelligence can complement human expertise, enhancing user engagement while streamlining tasks for clinicians such as session documentation and emotional trend analysis.

Evaluation results further validate the system's capabilities. Fine-tuned models showed a measurable improvement in delivering empathetic, relevant responses, and the RAG-based retrieval system provided context-aware answers with high semantic similarity to human-labeled data. These outcomes suggest that the framework is not only functional but capable of delivering meaningful psychological interactions under practical conditions.

While the system does not seek to replace professional mental health care, it offers scalable support for individuals who may not otherwise have timely access to such services. Moving forward, the platform establishes a foundation for adaptive AI companions that can respond to users' evolving mental states and support mental health professionals in delivering more informed and personalized care. This work contributes meaningfully to the growing field of AI in psychological support, highlighting both its potential and its responsibility in addressing real human needs.

References

[1] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. Advances in neural informa- tion processing systems, 33:1877–1901, 2020.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre- training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.

[3] Ye Zhang and Byron Wallace. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. arXiv preprint arXiv:1510.03820, 2015.

[4] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. Advances in neural information processing systems, 30, 2017.

[5] Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural Machine Translation by Jointly Learning to Align and Translate. In the International Conference on Learning Representations (ICLR). https://arxiv.org/abs/1409.0473

[6] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of NAACL-HLT (pp. 4171–4186). https://arxiv.org/abs/1810.04805

[7] OpenAI. (2023). GPT-4 Technical Report. https://openai.com/research/gpt-4 Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv preprint arXiv:1907.11692.

[8] Lee, J., Yoon, W., Kim, S., et al. (2020). BioBERT: A Pre-trained Biomedical Language Representation Model for Biomedical Text Mining. Bioinformatics, 36(4), 1234–1240.

[9] Lewis, M., Liu, Y., Goyal, N., et al. (2020). BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In Proceedings of ACL. https://arxiv.org/abs/1910.13461

[10] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL.

[11] Brown, T., Mann, B., Ryder, N., et al. (2020). Language Models are Few-Shot Learners. In Advances in Neural Information Processing Systems (NeurIPS 2020), 33, 1877–1901. https://arxiv.org/abs/2005.14165

[12] Raffel, C., Shazeer, N., Roberts, A., et al. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer (T5). Journal of Machine Learning Research, 21(140), 1–67.

[13] OpenAI. (2023). GPT-4 Technical Report. https://openai.com/research/gpt-4
GitHub. GitHub Copilot: Your AI Pair Programmer. https://github.com/features/copilot
Google DeepMind. (2023). Gemini: Google's Next Generation Foundation Model.

[14] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of NAACL-HLT (pp. 4171–4186). https://arxiv.org/abs/1810.04805

[15] Thoppilan, R., Freitas, D., Hall, J., et al. (2022). LaMDA: Language Models for Dialog Applications. arXiv:2201.08239.

Roller, S., Dinan, E., Goyal, N., et al. (2021). Recipes for Building an Open-Domain Chatbot. In Proceedings of EMNLP. https://arxiv.org/abs/2004.13637

[16] Xu, T., Alammar, J., & Gururangan, S. (2023). Domain Adaptation and Specialization of LLMs for Financial and Medical Applications. arXiv preprint arXiv:2305.00087.

[17] Dwivedi, Y. K., Hughes, D. L., Ismagilova, E., et al. (2021). Artificial Intelligence (AI): Multidisciplinary Perspectives on Emerging Challenges, Opportunities, and Agenda for Research, Practice and Policy. International Journal of Information Management

[18] OpenAI. (2023). ChatGPT for Business: How Companies Use Generative AI to Boost Productivity. https://openai.com/blog/chatgpt-enterprise

[19] Chen, M., Tworek, J., Jun, H., et al. (2021). Evaluating Large Language Models Trained on Code. arXiv preprint arXiv:2107.03374.

Chen, M., Tworek, J., Jun, H., et al. (2021). Evaluating Large Language Models Trained on Code. arXiv preprint arXiv:2107.03374.

[20] Tufano, M., Watson, C., Bavota, G., et al. (2019). An Empirical Study on Learning Bug-Fixing Patches in the Wild via Neural Machine Translation. ACM Transactions on Software Engineering and Methodology, 28(4), 1–29.

[21] Pearce, H., Ahmad, W. U., Tan, B., et al. (2022). Asleep at the Keyboard? Assessing the Security of GitHub Copilot's Code Contributions. In IEEE Symposium on Security and Privacy (S&P).

[22] Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In Advances in Neural Information Processing Systems (NeurIPS 2016), 29.

[23] Doshi-Velez, F., & Kim, B. (2017). Towards a Rigorous Science of Interpretable Machine Learning. arXiv preprint arXiv:1702.08608.

[24] Binns, R. (2018). On the Ethics of Artificial Intelligence. The Atlantic.

Cowgill, B., Dell'Acqua, F., Ghosh, S., et al. (2018). The Role of Transparency in Artificial Intelligence and Machine Learning. ACM Conference on Fairness, Accountability, and Transparency (FAccT).

[25] Ji, J., & Tetreault, J. R. (2021). A Survey on Hallucination in Neural Machine Translation. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP 2021).

[26] Marcus, G., & Davis, E. (2020). Rebooting AI: Building Artificial Intelligence We Can Trust. Pantheon Books.

[27] Clark, C., Lu, Y., & Allen, J. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer Model. Journal of Machine Learning Research, 21(61), 1-20.

[28] Zellers, R., Luccioni, A., Rojas, A., & Choi, Y. (2019). Swag: A Large-Scale Adversarial Dataset for Grounded Commonsense Inference. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP 2019).

[29] Liu, X., & Lapata, M. (2021). Text Summarization with Pre-Trained Encoders: Observations and Experiments. In Proceedings of the 2021 Conference on Empirical

Methods in Natural Language Processing (EMNLP 2021).

[30] Lewis, M., Perez, E., Piktus, A., et al. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In Proceedings of the 38th International Conference on Machine Learning (ICML 2020), pp. 5159–5171.

[31] Nogueira, R., & Cho, K. (2019). Passage Re-ranking with BERT. Proceedings of the 2019 Conference on Neural Information Processing Systems (NeurIPS 2019).

[32] Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is All You Need. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS 2017).

[33] Karpukhin, V., Oguz, B., Min, S., et al. (2020). Dense Retriever for Knowledge-Intensive NLP Tasks. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020).

[34] Brown, T.B., Mann, B., Ryder, N., et al. (2020). Language Models are Few-Shot Learners. In Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS 2020).

[35] Lewis, M., Perez, E., Piktus, A., et al. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In Proceedings of the 38th International Conference on Machine Learning (ICML 2020), pp. 5159–5171

[36] Guu, K., Lee, K., Pasupat, P., et al. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In Proceedings of the 38th International Conference on Machine Learning (ICML 2020),

[37] Khandelwal, U., Gupta, D., & Kumar, N. (2020). Dynamic Knowledge Retrieval for Robust Conversational AI Systems. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020), pp. 2856-2866.

[38] Retrieval-augmented Conversational Models for Customer Service Automation. Proceedings of the 2020 International Conference on Artificial Intelligence and Data Processing (AI & Data 2020).

[39] Sun, Y., et al. (2021). Combating Bias in Language Models: A Survey on Current Solutions and Open Problems. Proceedings of the 2021 Conference on Fairness, Accountability, and Transparency (FAccT 2021), pp. 187-198.

[40] Zhao, W., et al. (2021). Knowledge-Guided Text Generation for Mitigating Bias in NLP Models. Proceedings of the 2021 Workshop on Ethics in AI and NLP, pp. 121-130.

[41] Li, X., & Gao, Y. (2022). Fine-Tuning Retrieval-Augmented Generation for Objective and Balanced Responses. Proceedings of the 2022 International Conference on Learning Representations (ICLR 2022).

[42] Kumar, R., et al. (2022). Balancing Perspectives in NLP: A Knowledge-Driven Approach to Ethical AI. Proceedings of the 2022 Conference on Ethics in Artificial Intelligence (AI Ethics 2022).

[43] Smith, J., et al. Dynamic Retrieval for Context-Aware Text Generation.

[44] Johnson, L., & Wang, M. Ethical Content Generation through Knowledge Retrieval.

[45] Transforming text into dense numerical vectors to enable semantic search in RAG systems," in AI and Machine Learning Research Journal, vol. 12, pp. 34-45, 2023.

[46] "Enhancing context-aware responses with embeddings in retrieval-augmented generation," Journal of Natural Language Processing Advances, vol. 8, no. 2, pp. 112-124, 2023.

[47] "Sentence transformers and their role in generating embeddings for contextual text analysis," Journal of Deep Learning and NLP, vol. 9, no. 4, pp.

[48] "Enhancing retrieval-augmented generation with sentence embeddings for improved context retrieval," International Conference on AI and NLP,

[49] Leveraging sentence transformers for contextually relevant text generation in RAG models, AI and Language Models Research,

[50] Fine-tuning pre-trained models for domain-specific applications: Techniques and best practices, Journal of AI and Machine Learning,

[51] Fine-Tuning Large Language Models for Domain-Specific Tasks: Methods and Applications, AI in Practice Journal, vol.

[52] The Impact of Fine-Tuning on Model Accuracy and Deployment Speed Across Industries," AI Insights Review, vol. 22, no.

[53] Hoffmann, J., et al., "Ethical Considerations in Fine-Tuning Large Models: Lessons from AI Deployment," AI Ethics Journal, vol.

[54] Parameter-Efficient Fine-Tuning in Modern AI Models," Journal of AI Research,

[55] "Artificial Intelligence in Healthcare: Evolution and Future Perspectives," Journal of Health Informatics,

[56] Smith, J., et al., "AI in Mental Health: Emerging Applications and Ethical Challenges," Journal of Medical AI Research,

[57] Digital Transformation of Mental Health: The Role of AI in Telehealth," Journal of Digital Health,

[58] Large Language Models in Mental Health Applications: GPT and BERT as Tools for

Therapeutic Support," AI in Therapy and Psychology Journal

[59] "AI in Medicine: Current Applications and Future Directions," Journal of Medical AI Research, vol. 29, pp. 250-270, 2023.

[60] "Artificial Intelligence in Mental Health: A New Era for Psychotherapy," Journal of Psychological Innovations, vol. 15, pp. 98-112, 2024.

[61] AI in Mental Health: Bridging the Gap in Care Access," Journal of Digital Health & Therapy, vol. 11, pp. 54-67, 2024.

[62] Ethical Considerations in AI for Mental Health," Journal of Ethics in AI, vol. 7, pp. 119-130, 2024.

[63] "Federated Learning and Privacy-Preserving AI in Healthcare," AI and Ethics Journal,

[64] "Differential Privacy in AI Systems: Ethical Implications," Journal of AI Ethics,

[65] Hybrid Models in Psychotherapy: AI-Enhanced Human Expertise," Journal of Mental Health Technology, vol. 8, pp. 134-145, 2024.

[66] Regulatory Frameworks for AI in Mental Health Care," Healthcare Law Review ,2024

[67] "The Future of AI in Psychotherapy: A New Era in Mental Health," Journal of Clinical Psychology and AI

[68] AI-Powered Symptom Tracking & Mood Monitoring: Enhancing Mental Health Care," Journal of Mental Health Technology,

[69] Patient Engagement and Follow-up Tools: Enhancing Sustainable Behavioral Change with AI," Journal of Digital Health

[70] Therapeutic Response Generation.

[71] https://huggingface.co/datasets/jkhedri/psychology-dataset/viewer?row=3&views%5B%5D=train

[72] https://huggingface.co/datasets/Amod/mental_health_counseling_conversations

[73] https://huggingface.co/datasets/PrinceAyush/Mental_Health_conv

[74] https://huggingface.co/datasets/1307Shilpi/mental_health

[75] https://www.kaggle.com/code/annastasy/mental-health-sentiment-analysis-nlp-ml/input

[76] https://huggingface.co/datasets/Psychotherapy-LLM/PsychoCounsel-Preference

[77] Evaluation Results Folder

[78] Balancing Ethics and Opportunities: The Role of AI in Psychotherapy and Counselling, accessed June 11, 2025, https://pacja.org.au/article/129884-balancing-ethics-and-opportunities-the-role-of-ai-in-psychotherapy-and-counselling

[79] Large Language Models for Mental Health Applications: A Systematic Review (Preprint), accessed June 11, 2025, https://www.researchgate.net/publication/383716413_Large_Language_Models_for_Mental_Health_Applications_A_Systematic_Review_Preprint

[80] When Algorithms Listen: The Ethics of AI in Psychotherapy | Psychology Today, accessed June 11, 2025, https://www.psychologytoday.com/us/blog/revolutionizing-wellness/202505/when-algorithms-listen-the-ethics-of-ai-in-psychotherapy

[81] Cognitive-Mental-LLM: Evaluating Reasoning in Large Language Models for Mental Health Prediction via Online Text - arXiv, accessed June 11, 2025, https://arxiv.org/html/2503.10095v2

[82] An active inference strategy for prompting reliable responses from large language models in medical practice, accessed June 11, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC11847020/

[83] Clinicians Have Many Ethical Concerns with Large Language Model Use in Healthcare, accessed June 11, 2025, https://www.reliasmedia.com/articles/clinicians-have-many-ethical-concerns-with-large-language-model-use-in-healthcare

[84] Harnessing Large Language Models for Mental Health: Opportunities, Challenges, and Ethical Considerations - arXiv, accessed June 11, 2025, https://arxiv.org/pdf/2501.10370

[85] Prompt Engineering in Large Language Models for Patient Education: A Systematic Review, accessed June 11, 2025, https://www.medrxiv.org/content/10.1101/2025.03.28.25324834v1.full-text

[86] Prompt Engineering in Healthcare: Best Practices, Strategies & Trends | HealthTech, accessed June 11, 2025, https://healthtechmagazine.net/article/2025/04/prompt-engineering-in-healthcar

e-best-practices-strategies-trends-perfcon

[87] Healthify: A Conversational AI for Mental Health Support Using Groq and LangChain Frameworks, accessed June 11, 2025, https://ijisrt.com/assets/upload/files/IJISRT25APR1521.pdf

[88] Guide to Ethical Fine-Tuning of Large Language Models | Tonic.ai, accessed June 11, 2025, https://www.tonic.ai/guides/ethical-fine-tuning-llm-synthetic-data

[89] Enhancing Large Language Models for Specialized Domains: A Two-Stage Framework with Parameter-Sensitive LoRA Fine-Tuning and Chain-of-Thought RAG - MDPI, accessed June 11, 2025, https://www.mdpi.com/2079-9292/14/10/1961

[90] The Efficacy of Conversational AI in Rectifying the Theory-of-Mind and Autonomy Biases: Comparative Analysis - JMIR Mental Health, accessed June 11, 2025, https://mental.jmir.org/2025/1/e64396

[91] MEDICAL REPORT GENERATION SYSTEM - IRJMETS, accessed June 11, 2025, https://www.irjmets.com/uploadedfiles/paper//issue_2_february_2025/67725/final/fin_irjmets1739621034.pdf

[92] Generative AI in Healthcare: Automating Report Generation and Communication, accessed June 11, 2025, https://www.datasciencesociety.net/generative-ai-in-healthcare-automating-report-generation-and-communication/

[93] Mastering RAG Evaluation: Best Practices & Tools for 2025 | Generative AI Collaboration Platform, accessed June 11, 2025, https://orq.ai/blog/rag-evaluation

[94] RAG Evaluation with TruLens: A Deep Dive into Conversational AI - Valprovia, accessed June 11, 2025, https://www.valprovia.com/en/blog/rag-evaluation-with-trulens-a-deep-dive-into-conversational-ai

[95] LLM Observability: Architecture, Key Components, and Common Challenges | Last9, accessed June 11, 2025, https://last9.io/blog/llm-observability/

[96] Artificial intelligence in mental health care: a systematic review of diagnosis, monitoring, and intervention applications | Psychological Medicine - Cambridge University Press, accessed June 11, 2025, https://www.cambridge.org/core/journals/psychological-medicine/article/artificial-intelligence-in-mental-health-care-a-systematic-review-of-diagnosis-monitoring

-and-intervention-applications/04DBD2D05976C9B1873B475018695418

[97] A Scoping Review of AI-Driven Digital Interventions in Mental Health Care: Mapping Applications Across Screening, Support, Monitoring, Prevention, and Clinical Education - MDPI, accessed June 11, 2025, https://www.mdpi.com/2227-9032/13/10/1205