Faculty of Engineering and Information Technology

Mahmoud Essam Fathy 191900203

Hader Farag 191900162

Mariem Kamel 191900053

Mohamed Khaled 191900260

Alaa Kamal 191900057

Graduation Project Report

Wanas

December 2023

Under SuperVision of Dr. Amr Zamil

Abstract

This graduation project delves into the realm of mental health support by introducing an innovative application designed to act as a virtual therapist. In a world where the pace of life is relentless and mental well-being is of paramount importance, our project aims to leverage artificial intelligence and machine learning to provide accessible and personalized assistance. The virtual therapist application offers users a confidential and empathetic space to express their thoughts and emotions, fostering a sense of connection and understanding. This documentation explores the design principles, underlying technologies, and ethical considerations that guided the development of the application. Additionally, it delves into the potential impact of such a solution on mental health care, addressing the pressing need for accessible and stigma-free support in today's digital landscape. Join us on this journey as we navigate the intersection of technology and mental health, contributing to the ongoing dialogue about innovative solutions for well-being in the modern age.

Acknowledgements

Contents

_____

Chapter 1

Introduction

Mental health is a state of well-being that allows an individual to feel, think and act to their full potential. They are an integral part of health and well-being that underpin our individual and collective abilities to make decisions, establish relationships and shape the world in which we live. Mental health affects every aspect of our lives, including our physical health, our social relationships, our productivity at work or school, and our general sense of well-being. When our mental health is good, we are able to deal with daily challenges and feel happy and satisfied with life. However, mental health is an important issue that is often overlooked. In fact, according to the World Health Organization, about 1 in 4 people in the world will suffer from a mental disorder at some point in their lives. There are many reasons why people hesitate to seek help from psychiatrists. One reason is fear of stigma. There are still a lot of social judgments around mental health, and some people fear they will be seen as weak or unstable if they seek help. Another reason for hesitation is fear of cost. Mental health care can be expensive, especially in low- or middle-income countries. Finally, some people may find it difficult to talk about their feelings and experiences with a psychologist. They may feel ashamed or afraid of being judged. Therefore, we present the WANAS application, through which we provide an innovative solution to these challenges, where individuals can talk to artificial intelligence to obtain support and guidance at any time.

1.1   ProjectIdea

The graduation project centers around the development of an application designed to function as a virtual therapist. The primary goal is to provide accessible and personalized mental health support in an increasingly fast-paced and interconnected world. The application utilizes artificial intelligence and machine learning to create an empathetic and intelligent virtual therapist capable of engaging with users in a meaningful way.

Key Features:

Personalized Support: The virtual therapist is designed to offer tailored assistance, taking into account individual needs, preferences, and mental health concerns. Through advanced algorithms, the application adapts its responses to the user's specific emotional state and context.

Confidential and Non-Judgmental Space: Recognizing the importance of privacy in mental health discussions, the application creates a secure and confidential environment. Users can freely express their thoughts and emotions without fear of judgment, fostering a sense of trust and openness.

User-Friendly Interface: The project prioritizes a user-friendly interface to ensure that individuals, regardless of their technological proficiency, can easily navigate and interact with the virtual therapist. The goal is to make mental health support accessible to a broad audience.

Integration of AI and Machine Learning: Leveraging the latest advancements in artificial intelligence and machine learning, the virtual therapist continuously learns and improves its responses over time. This dynamic adaptation ensures a more effective and personalized support system for users.

Addressing Mental Health Stigma: The project aims to contribute to the destigmatization of mental health issues by providing a digital platform that encourages users to seek help in a discreet and comfortable manner.

Throughout the development process, ethical considerations are a key focus, ensuring user privacy, data security, and adherence to established mental health care guidelines. The project envisions making a meaningful impact on the mental well-being of individuals globally by offering a modern, technology-driven approach to mental health support.

1.1.1   why?

The need for the virtual therapist application arises from several compelling reasons, addressing significant challenges in the current landscape of mental health care:

Accessibility and Timeliness: Traditional mental health services often face challenges related to accessibility and timely support. The virtual therapist application fills this gap by providing immediate assistance, allowing users to access support whenever they need it, without the constraints of scheduling appointments or geographical limitations.

Stigma Reduction: Mental health stigma remains a formidable barrier to seeking help. By offering a confidential and non-judgmental space, the application contributes to destigmatizing mental health issues. Users can engage with the virtual therapist discreetly, reducing the fear of social repercussions and promoting a more open dialogue about mental health.

Global Reach: In a world where mental health issues are pervasive and diverse, a digital solution provides a scalable and globally accessible platform. The virtual therapist application has the potential to reach individuals in remote areas or regions with limited mental health resources, democratizing access to support services.

Personalization and Adaptability: Every individual's mental health journey is unique, requiring personalized support. The application utilizes artificial intelligence and machine learning to adapt its responses to the specific needs and emotional states of users. This personalized approach enhances the effectiveness of the support provided.

24/7 Support: Mental health concerns do not adhere to a fixed schedule. The virtual therapist application offers continuous support, 24 hours a day, 7 days a week. This ensures that users can seek assistance at any time, particularly during moments of crisis or when faced with urgent emotional challenges.

Complementary Support to Traditional Services: The application is designed to complement traditional mental health services rather than replace them. It serves as an additional resource, offering immediate support and guidance, while also encouraging users to seek professional help when necessary.

Technological Advancements: With the rapid advancement of technology, leveraging artificial intelligence and machine learning in mental health care represents a logical progression. The

virtual therapist application harnesses these technologies to enhance the quality and accessibility of mental health support.

In summary, the virtual therapist application is a response to the evolving needs of individuals in a digital age, providing a convenient, stigma-free, and personalized solution to the challenges associated with mental health care. It aims to empower users to take control of their mental well-being by offering support that is accessible, adaptable, and considerate of the diverse nature of mental health experiences.

Chapter 2

Background

2.1 AI

Artificial Intelligence (AI) refers to the development of computer systems that can perform tasks that typically require human intelligence. These tasks include learning from experience, understanding natural language, recognizing patterns, solving problems, and making decisions. AI systems can be categorized into two main types: narrow or weak AI, which is designed for a specific task, and general or strong AI, which possesses human-like cognitive abilities across a range of tasks.

2.2 machine-learning

Machine Learning (ML) is a subfield of artificial intelligence (AI) that focuses on the development of algorithms and statistical models that enable computer systems to perform tasks without being explicitly programmed. The essence of machine learning lies in the ability of machines to learn from data, recognize patterns, and make decisions or predictions based on that learning.

2.3 Neural Network

A neural network is a computational model inspired by the structure and functioning of the human brain. It is a fundamental component of deep learning, a subset of machine learning. Neural networks consist of interconnected nodes, or artificial neurons, organized into layers. The

three main types of layers in a neural network are the input layer, hidden layers, and output layer.

2.3.1   Nodes

In a neural network, nodes, also known as neurons or units, are fundamental building blocks that process and transmit information. Nodes are organized into layers, and each layer serves a specific function in the network. The three main types of layers in a neural network are the input layer, hidden layers, and output layer.

- Input Layer: The input layer is the initial layer that receives raw data or features. Each node in the input layer represents a feature or attribute of the input data. The number of nodes in the input layer is determined by the dimensionality of the input data.

- Hidden Layers: Hidden layers are intermediate layers between the input and output layers. Each node in a hidden layer processes information and contributes to the network's ability to learn complex patterns. The term "hidden" arises because these layers are not directly observable in the input or output data.

- Output Layer: The output layer produces the final results or predictions of the neural network. The number of nodes in the output layer depends on the nature of the task. For example, in a binary [1]classification task, there might be one node for each class, while in a regression task, there might be a single node.

Each connection between nodes is associated with a weight, which represents the strength of the connection. During training, these weights are adjusted to enable the network to learn from the input data and make accurate predictions.

2.3.2   Activation Function

An activation function is a crucial component in artificial neural networks and deep learning models. It introduces non-linearity to the network, allowing it to learn complex relationships in the data. The activation function operates on the input received by a node (or neuron) and produces an output that is used as the input for the next layer of the network.

Here are some commonly used activation functions in neural networks:

- Sigmoid Function (Logistic Activation):

Formula:

$$f(x) = \frac{1}{1 + e^{-x}} \tag{2.1}$$

Range: $(0, 1)$

Used in the output layer of binary classification models to squash the output into a probability range.

- Hyperbolic Tangent Function (tanh):

  Formula:

$$f(x) = \frac{e^{2x} - 1}{e^{2x} + 1} \tag{2.2}$$

  Range: (-1, 1)

  Similar to the sigmoid function but with a broader output range. It is often used in hidden layers.

- Rectified Linear Unit (ReLU):

  Formula:

$$f(x) = \max(0, x) \tag{2.3}$$

  Range: $[0, \infty)[0, \infty)$

  Widely used in hidden layers due to its simplicity and effectiveness in training deep networks. However, it can suffer from the "dying ReLU" problem when neurons become inactive during training.

- Leaky ReLU:

  Formula:

$$f(x) = \max(\alpha x, x) \tag{2.4}$$

  Introduces a small slope for negative values, addressing the dying ReLU problem.

- Parametric ReLU (PReLU):

  Similar to Leaky ReLU but allows the slope to be learned during training rather than being a fixed parameter.

- Exponential Linear Unit (ELU):

Formula:

$$\text{ELU}(x) = \begin{cases} x & \text{if } x \geq 0 \\ \alpha(e^x - 1) & \text{if } x < 0 \end{cases} \tag{2.5}$$

A smooth alternative to ReLU that includes a saturation region for negative values.

- Softmax Function:

  Formula:

  $$\text{softmax}(x)_i = \frac{e^{x_i}}{\sum_j e^{x_j}} \tag{2.6}$$

  Converts a vector of real numbers into a probability distribution. Often used in the output layer for multi-class classification problems.

2.4 Deep-learning

Deep learning is a subset of machine learning that involves the use of neural networks with multiple layers (deep neural networks) to model and solve complex problems. The term "deep" refers to the depth of the neural network, which is characterized by having multiple layers, often referred to as hidden layers, between the input and output layers.

2.5 Natural Language Processing

[4] Natural Language Processing (NLP) is a field of artificial intelligence (AI) that focuses on the interaction between computers and human languages. The goal of NLP is to enable machines to understand, interpret, and generate human-like text. This interdisciplinary field involves aspects of computer science, linguistics, and cognitive psychology.

2.6 Language-Model

A language model is a type of artificial intelligence (AI) model designed to understand and generate human-like text based on the input it receives. These models are trained on vast amounts of textual data and can be used for a variety of natural language processing (NLP) tasks.

2.7   Large-Language-Model

A large language model typically refers to a sophisticated artificial intelligence (AI) model designed for natural language processing (NLP) tasks and capable of handling vast amounts of textual data. These models have a high number of parameters, allowing them to capture intricate patterns and relationships in language. The size of a language model is often measured in terms of the number of parameters, which can range from millions to billions.

2.8   Transformers in Deep Learning

Transformers are a crucial architectural innovation in deep learning that revolutionized natural language processing and other sequential data tasks. Introduced in the paper [7] "Attention is All You Need" by Vaswani et al., transformers have become the foundation for various state-of-the-art models.

Key Components

1. **Self-Attention Mechanism:** Allows the model to focus on different parts of the input sequence while processing each element.

2. **Multi-Head Attention:** Multiple self-attention heads run in parallel, capturing diverse aspects of relationships within the sequence.

3. **Positional Encoding:** Incorporates information about the position of tokens in the input sequence, as transformers do not inherently understand the sequential order of data.

4. **Feedforward Networks:** Non-linear transformations applied to the outputs of the attention mechanism.

Advantages

1. **Parallelization:** Transformers enable parallel processing of input sequences, making them computationally efficient.

2. **Capturing Long-Range Dependencies:** Self-attention allows transformers to capture relationships between distant elements in a sequence.

3. **Versatility:** Initially designed for natural language processing, transformers have proven

adaptable to various tasks, including image recognition and machine translation.

Applications

1. **BERT (Bidirectional Encoder Representations from Transformers):** Introduced bidirectional context understanding for natural language understanding tasks.

2. **GPT (Generative Pre-trained Transformer):** Utilizes transformers for autoregressive language modeling, generating coherent and contextually relevant text.

3. **Image Transformers:** Applied to computer vision tasks, demonstrating competitive performance in image classification.

Limitations

1. **Computational Complexity:** Transformers may be computationally expensive, particularly for large models.

2. **Sequential Structure Understanding:** While transformers excel at parallel processing, they lack inherent understanding of sequential order.

2.9   BERT Model

[3] BERT stands for Bidirectional Encoder Representations from Transformers and is a language representation model by Google. This pre-trained model can be fine-tuned for a variety of final tasks that might not be similar to the task model was trained on. BERT makes use of Transformer, an attention mechanism that learns contextual relations between words (or sub-words) in a text. In its vanilla form, Transformer includes two separate mechanisms — an encoder that reads the text input and a decoder that produces a prediction for the task. Since BERT's goal is to generate a language model, only the encoder mechanism is necessary. Historically, language models could only read text input sequentially – either left-to-right or right-to-left – but couldn't do both at the same time. BERT is different because it is designed to read in both directions at once. This capability, enabled by the introduction of Transformers, is known as bidirectionality.

2.9.1   Model architecture

L = Number of layers (i.e., #Transformer encoder blocks in the stack).

H = Hidden size (i.e. the size of q, k and v vectors).

A = Number of attention heads.

BERT Base: L=12, H=768, A=12.

Total Parameters=110M!

BERT Large: L=24, H=1024, A=16.

Total Parameters=340M!!

2.9.2   Pre-training BERT

The BERT model is trained on the following two unsupervised tasks.

- Masked Language Model (MLM)

  This task enables the deep bidirectional learning aspect of the model. In this task, some percentage of the input tokens are masked (Replaced with [MASK] token) at random and the model tries to predict these masked tokens — not the entire input sequence. The predicted tokens from the model are then fed into an output softmax over the vocabulary to get the final output words.

- Next Sentence Prediction (NSP)

  It is a technique used in BERT to help the model understand the relationship between two consecutive sentences. In simple terms, NSP is designed to teach BERT how sentences are connected and whether they logically follow each other.

2.10   Fine tuning

[5] Fine tuning is a machine-learning technique that involves making small changes to a pre-trained model to improve its performance on a specific task and helps the model better understand the specific context and language patterns of the task it is being fine-tuned for. This is more efficient and often yields better results than training a model from scratch, as the model already has a good understanding of the world and can use this knowledge to learn the new task more quickly. Suppose a healthcare organization wants to use LLM to assist doctors in generating patient reports from textual notes. While this model can understand and create gen-

eral text, it might not be optimized for intricate medical terms and specific healthcare jargon, So to enhance its performance for this specialized role, the organization fine-tunes the model on a dataset filled with medical reports and patient notes. And after fine-tuning, the model is primed to assist doctors in generating accurate and coherent patient reports, demonstrating its adaptability for specific tasks. If we didn't use this approach, we would initially have random weights. In that case, a complex architecture would require a lot of time for training. So briefly, fine-tuning refers to using the weights of an already trained network as the starting values for training a new network.

2.10.1   Why Fine-Tuning?

- **Transfer Learning:** Pre-trained models have already learned useful features from large datasets. Fine-tuning allows us to use this knowledge and apply it to a new related task. It's like taking knowledge gained in one area and using it in another.

- **Efficiency:** Training a model from scratch can be computationally expensive and time-consuming. Fine-tuning is more efficient since it starts with a model that has already learned a lot and then refines it for a specific task.

- **Limited Data:** In many real-world scenarios, collecting large amounts of labelled data for training a model from scratch might be impractical. Fine-tuning helps when you have limited data because the pre-trained model already has a good understanding of the general features or language meanings.

- **Domain-Specific Tasks:** Models trained on general tasks can be fine-tuned for more specific tasks in a particular domain. For example, a model trained on general text can be fine-tuned for sentiment analysis in customer reviews.

2.10.2   Capabilities of an LLM after Fine tuning

- Sentiment Analysis: Boost the model performance in understanding and categorizing emotions and opinions in textual content.

- Named Entity Recognition (NER): For tasks that involve pinpointing the names of individuals, organizations, locations, and various entities, fine-tuning the model can significantly enhance its precision and accuracy in recognizing and classifying these key elements.
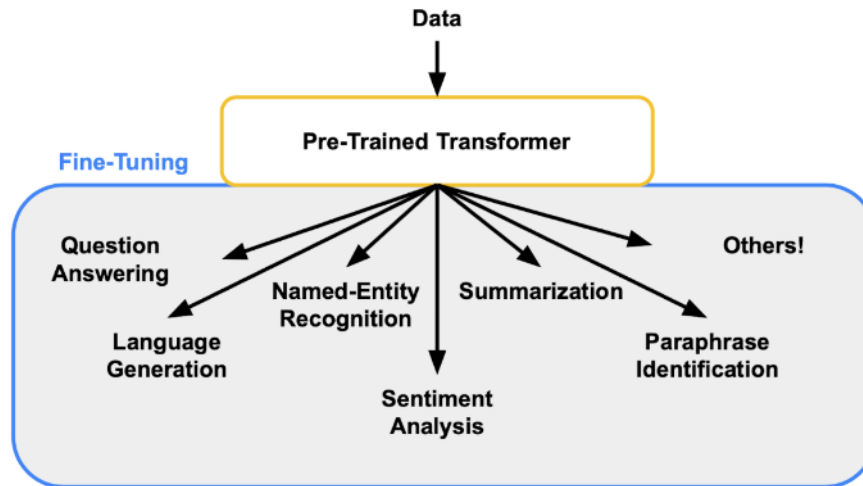
- Text Generation: Fine-tuning is a powerful tool for customizing the model's text generation abilities. This process allows you to make the model's output to follow to specific writing styles, tones, or themes.

- Question Answering: When dealing with tasks that require answering questions based on contextual information we may using fine tuning. It equips the model with the ability to understand and extract relevant details, enabling it to provide accurate and context-aware responses.

2.10.3   Supervised fine-tuning:

Supervised fine-tuning is a process in machine learning which we are using a pre-trained model, and trying to adapted it for a specific task using labelled data. In this approach, the model's parameters are adjusted based on a clear set of input-output pairs, where the output is known and provided during the training phase. The objective is for the model to learn the intricate patterns and relationships within the labelled data, allowing it to generalize its understanding to new, similar examples and enhancing its performance in making accurate predictions or classifications.

The approach involves the following steps:

1. Choose the fine-tuning task, such as text generation or analysis.

2. Prepare datasets containing questions and answers related to the chosen topic. Specify prompts for generating questions and completing tasks.

3. Select the base model, e.g., "arabert."

4. Fine-tune the model using a supervised approach, specifying relevant parameters and training methods.

5. Evaluate the performance of the fine-tuned model using appropriate metrics.

### 2.10.4 Reinforcement Learning Overview

[2] Reinforcement Learning (RL) is a machine learning training method that rewards desired behaviors and penalizes undesired ones. It involves an agent perceiving and interpreting its environment, taking actions, and learning through trial and error. Key components include the agent, environment, observation, state, action, reward, and policy.

How Reinforcement Learning Works

RL involves an agent exploring an unknown environment to achieve a goal, maximizing expected cumulative reward. The main elements include the agent, environment, policy, and reward signal. The value function captures the 'goodness' of a state, and the Bellman equation is a key concept in RL, used for calculating value functions.

Approaches to Implement RL Algorithms

Three approaches:
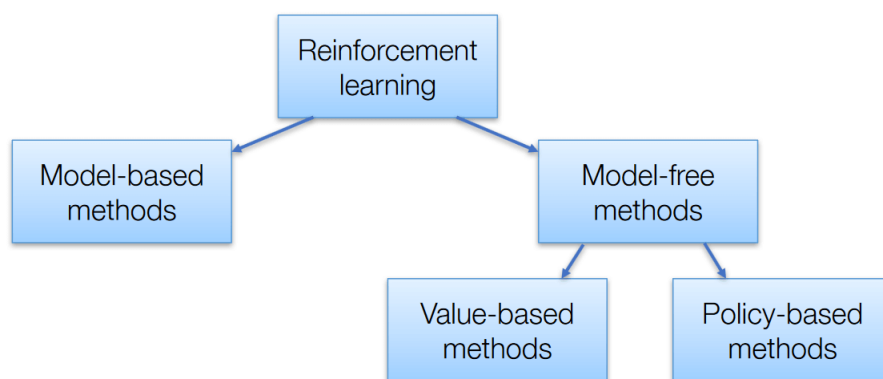
1. **Value-Based:** Maximizing a value function.

2. **Policy-Based:** Developing a strategy for maximum future rewards.

3. **Model-Based:** Creating a virtual model to assist learning in a specific environment.

Bellman Equation

Utilizes the action, state, reward, discount factor, and value of the starting point to estimate values and maximize rewards. The key-elements used in Bellman equations are:

- Action performed by the agent is referred to as "a"

- State occurred by performing the action is "s."

- The reward/feedback obtained for each good and bad action is "R."

- A discount factor is Gamma "γ."

- The value of your starting point is the reward you expect to get from being there, plus the value of wherever you land next.

- On-Policy Value Functions The Bellman equation for the on-policy value functions is given by:

$$V^\pi(s) = \sum_a \pi(a|s) \sum_{s',r} p(s', r|s, a) \left[r + \gamma V^\pi(s') \right] \tag{2.7}$$

- The corresponding update rule for the action-value function is:

$$Q^\pi(s, a) = \sum_{s',r} p(s', r|s, a) \left[r + \gamma \sum_{a'} \pi(a'|s') Q^\pi(s', a') \right] \tag{2.8}$$

- Optimal Value Functions

  The Bellman equation for the optimal value functions is given by:

$$V^*(s) = \max_a \sum_{s',r} p(s', r|s, a) \left[r + \gamma V^*(s') \right] \tag{2.9}$$

- The corresponding update rule for the optimal action-value function is:

$$Q^*(s, a) = \sum_{s',r} p(s', r|s, a) \left[r + \gamma \max_{a'} Q^*(s', a') \right] \tag{2.10}$$

Q-Learning

A common RL algorithm that approximates the optimal action-value function. Q-learning methods learn an approximator for the optimal action-value function, typically using an objective
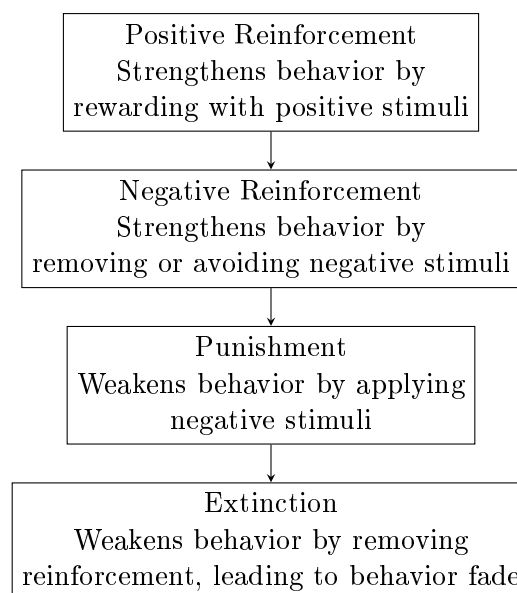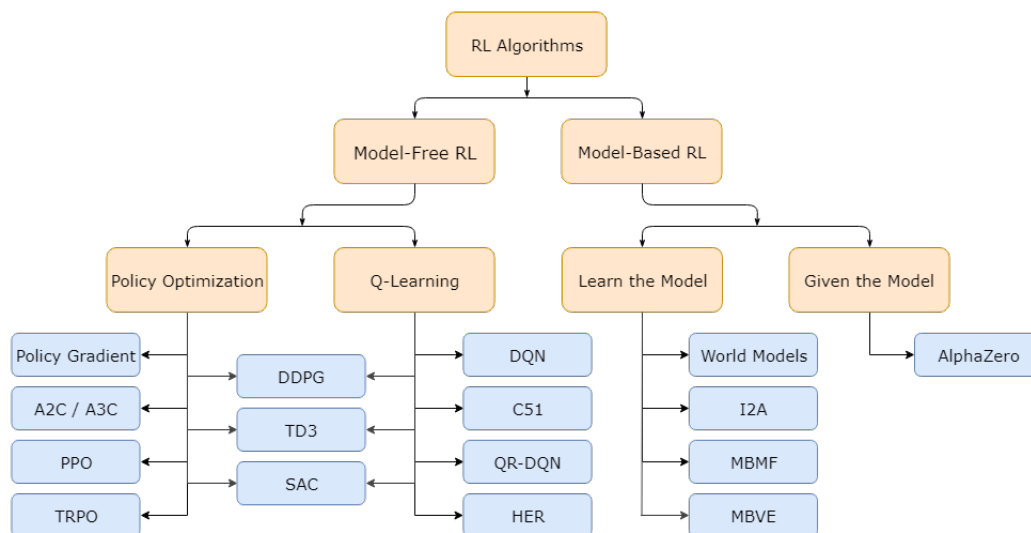
Figure 2.1: Types of Reinforcement Learning

function based on the Bellman equation.

$$Q(s,a) = \sum_{s',r} p(s',r \mid s,a) \left[r + \gamma \max_{a'} Q(s',a') \right] \tag{2.11}$$

Types of Reinforcement Learning

1. **Positive Reinforcement:** Strengthens and increases behavior.

2. **Negative Reinforcement:** Strengthens behavior by avoiding negative conditions.

17

Reinforcement Learning Applications

- **Robotics:** Navigation, Robo-soccer, etc.

- **Control:** Adaptive control in factory processes, telecommunication, etc.

- **Game Playing:** Tic-tac-toe, chess, etc.

- **Chemistry:** Optimizing chemical reactions.

- **Business:** Strategy planning.

- **Manufacturing:** Robotic tasks in various industries.

- **Finance Sector:** Evaluating trading strategies.

2.10.5   What is wanas?

Wanas is conceptualized as a therapist that acts as a supportive friend to help individuals navigate daily mental challenges. The approach involves designing a model where different sub-models handle specific therapeutic tasks, and a larger model decides which sub-model to deploy at different times. The design process drew inspiration from interviews with therapists to understand how they communicate with patients, aiming to create interactions that bring happiness and a positive outlook to individuals' lives. This approach seeks to emulate the empathetic and uplifting qualities of human therapeutic interactions in an AI-driven model.

How therapist function?

- **Active Listening:** Therapists practice active listening, fully concentrating, understanding, responding, and remembering what the client is saying.

- **Empathy:** Showing empathy involves understanding and sharing the feelings of the client, expressing genuine understanding through verbal and non-verbal cues.

- **Open-ended Questions:** Therapists ask open-ended questions to encourage clients to share more about their thoughts and feelings.

- **Reflective Responses:** Therapists use reflective responses, paraphrasing or summarizing what the client has said to demonstrate active listening.

- **Clarification:** Seeking clarification when there is ambiguity or when more information

is needed to ensure a shared understanding.

- **Non-verbal Communication:** Paying attention to non-verbal cues such as body language, facial expressions, and tone of voice.

- **Validation:** Validating the client's experiences and emotions without judgment to build a trusting relationship.

- **Encouragement:** Offering encouragement to motivate clients to explore difficult topics and recognize their strengths and progress.

- **Cultural Sensitivity:** Being mindful of cultural differences and adapting communication styles to be sensitive to the client's cultural background.

- **Respect and Non-judgment:** Communicating with respect and without judgment to create a safe space for clients to share openly.

- **Feedback:** Providing constructive feedback in a supportive and collaborative manner to help clients gain insights into their behavior and thought patterns.

- **Goal Setting:** Collaboratively setting therapeutic goals to ensure focused and aligned therapy that empowers clients to take an active role in their own growth.

- **Boundaries:** Communicating and maintaining clear professional boundaries to ensure a safe and ethical therapeutic relationship.

- **Trauma-Informed Communication:** Using trauma-informed communication for clients who have experienced trauma, prioritizing safety and avoiding retraumatization.

2.10.6   AraBert

[6] Arabic BERT," is a pre-trained language model designed specifically for the Arabic language. BERT, or Bidirectional Encoder Representations from Transformers, is a transformer-based model developed by Google that has proven to be highly effective in capturing context and semantics in natural language.

AraBERT, as an adaptation of BERT for Arabic, has been trained on a large corpus of Arabic text to understand the nuances and intricacies of the language. This pre-training allows AraBERT to generate high-quality contextualized embeddings for Arabic words and sentences. These embeddings can then be fine-tuned for specific downstream tasks, such as sentiment

analysis, named entity recognition, or other natural language processing tasks in the Arabic language domain.

In summary, AraBERT is a powerful tool for natural language processing in Arabic, leveraging the strengths of BERT's architecture while specifically tailored for the characteristics of the Arabic language. which is traied on many datasets such as

2.10.7   Our Goal

The goal is to create multiple specialized models for distinct tasks, allowing the system to choose the most appropriate model based on patient responses. This involves fine-tuning a Language Model (LLM), such as AraBert, which has been pre-trained on extensive Arabic datasets. The pre-training equips the model with a comprehensive understanding of Arabic language rules, enabling it to effectively handle various language-related tasks in a healthcare or therapeutic context.

_____

Chapter 3

Analysis and Design

3.1 Diagrams

3.1.1 Activity Diagram
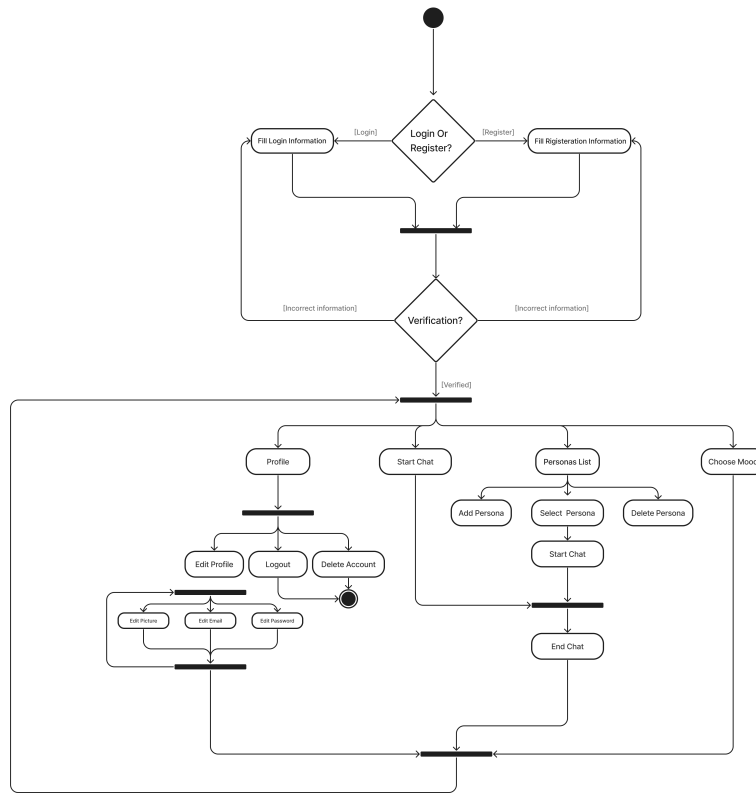
## 3.1.2   User Flow
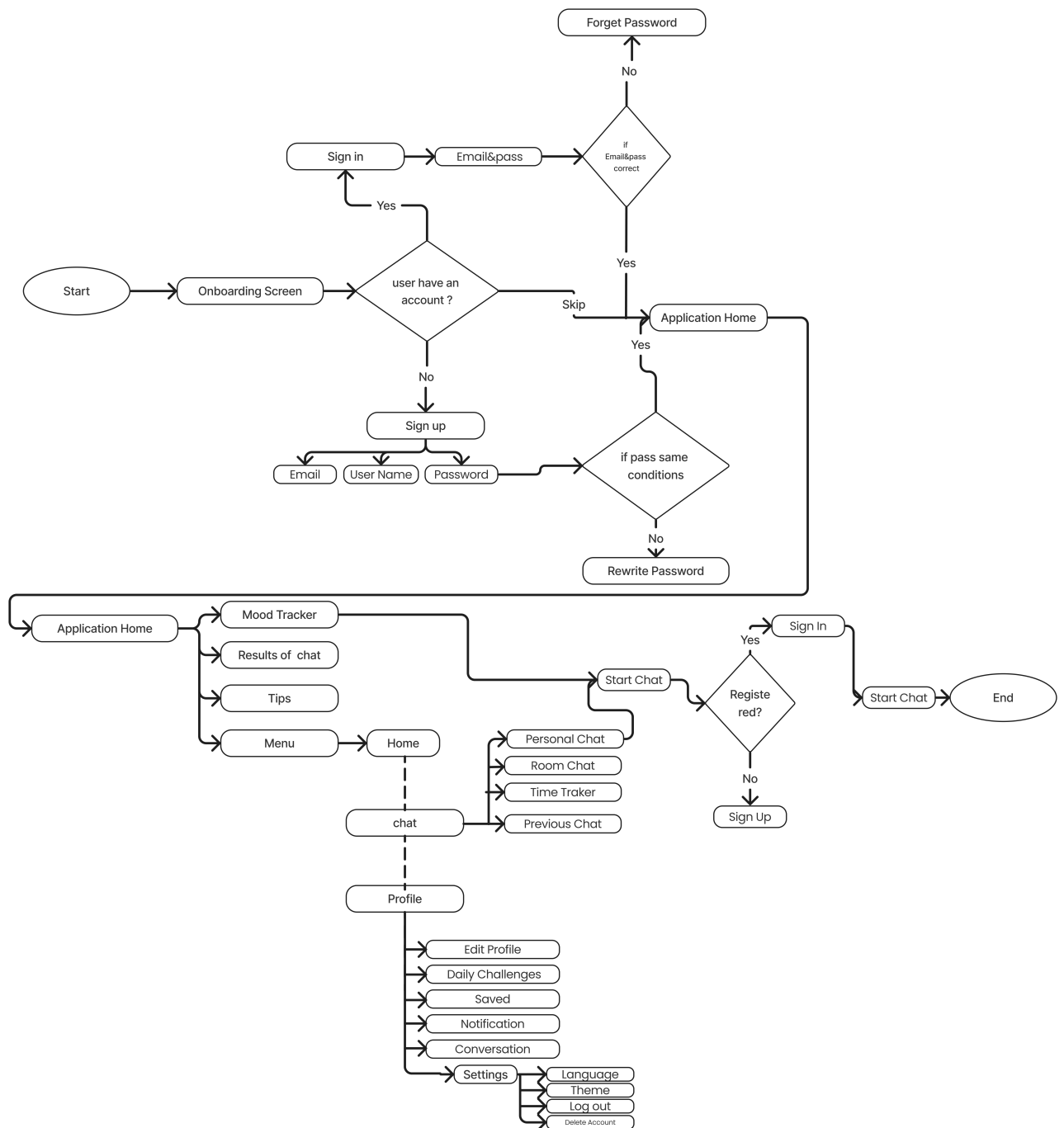
### 3.1.3   Block Diagram
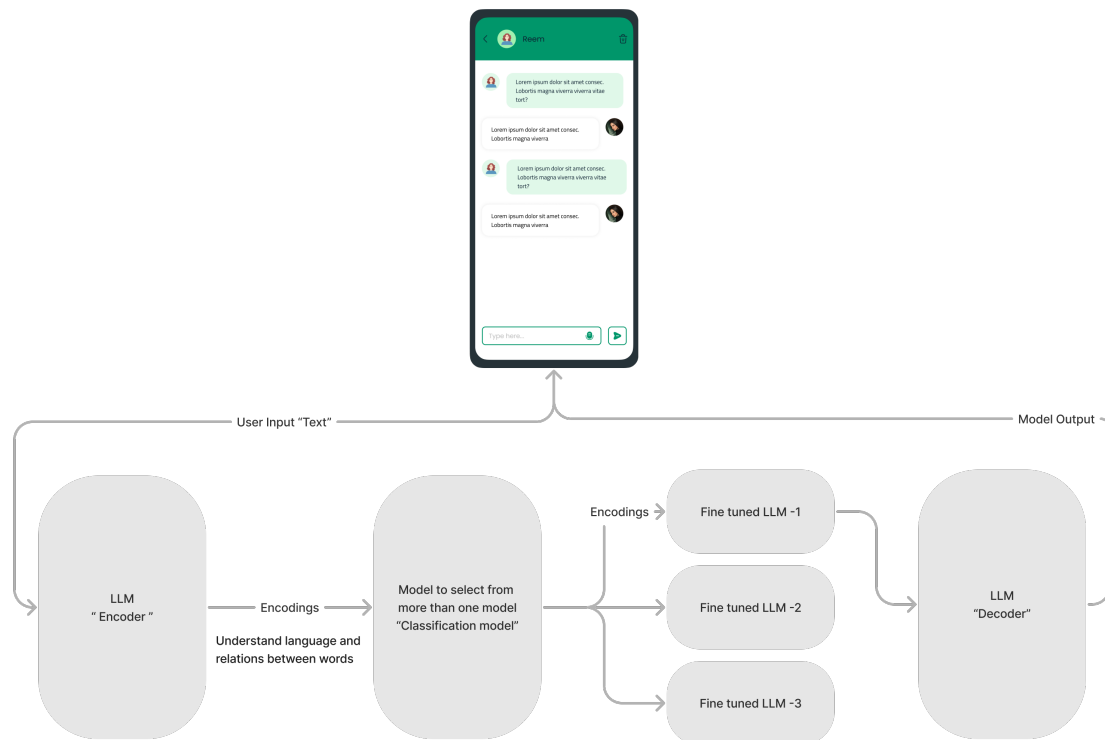
3.1.4   Class Diagram
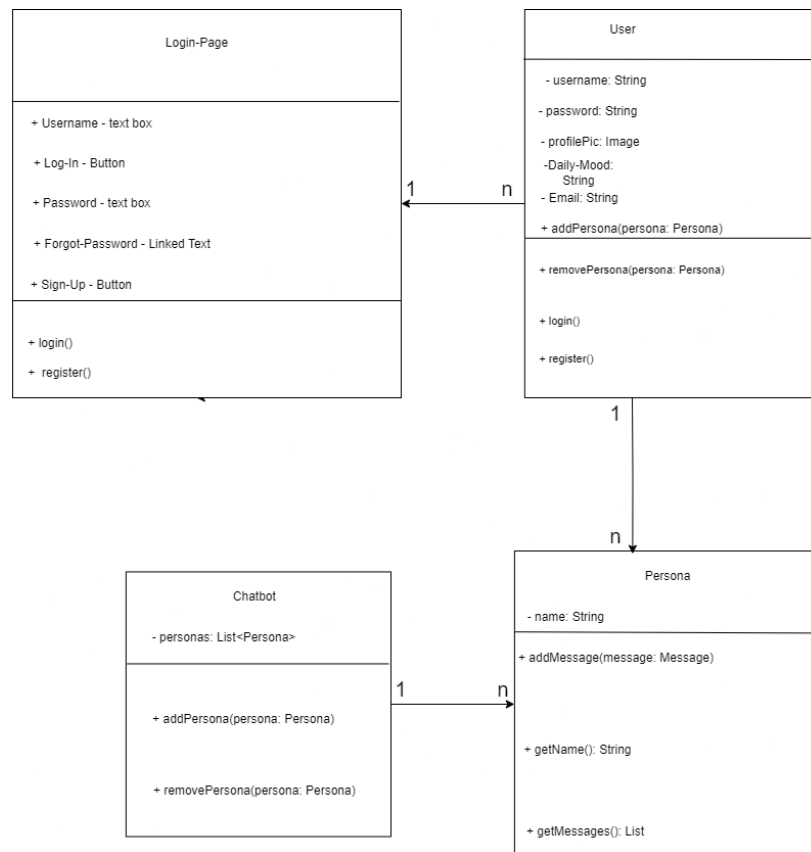
3.1.5　Sequence Diagram
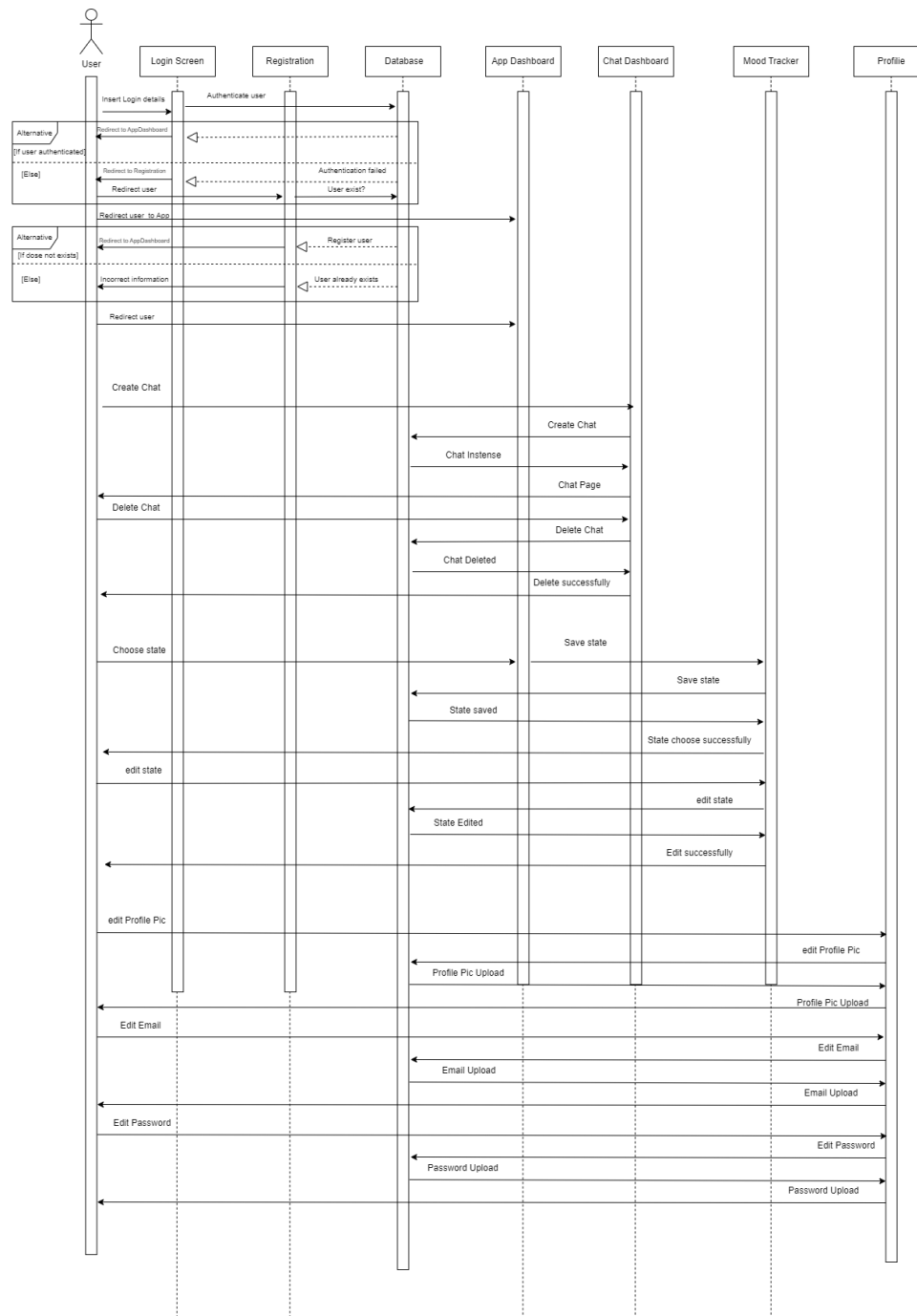
3.1.6   Use Case
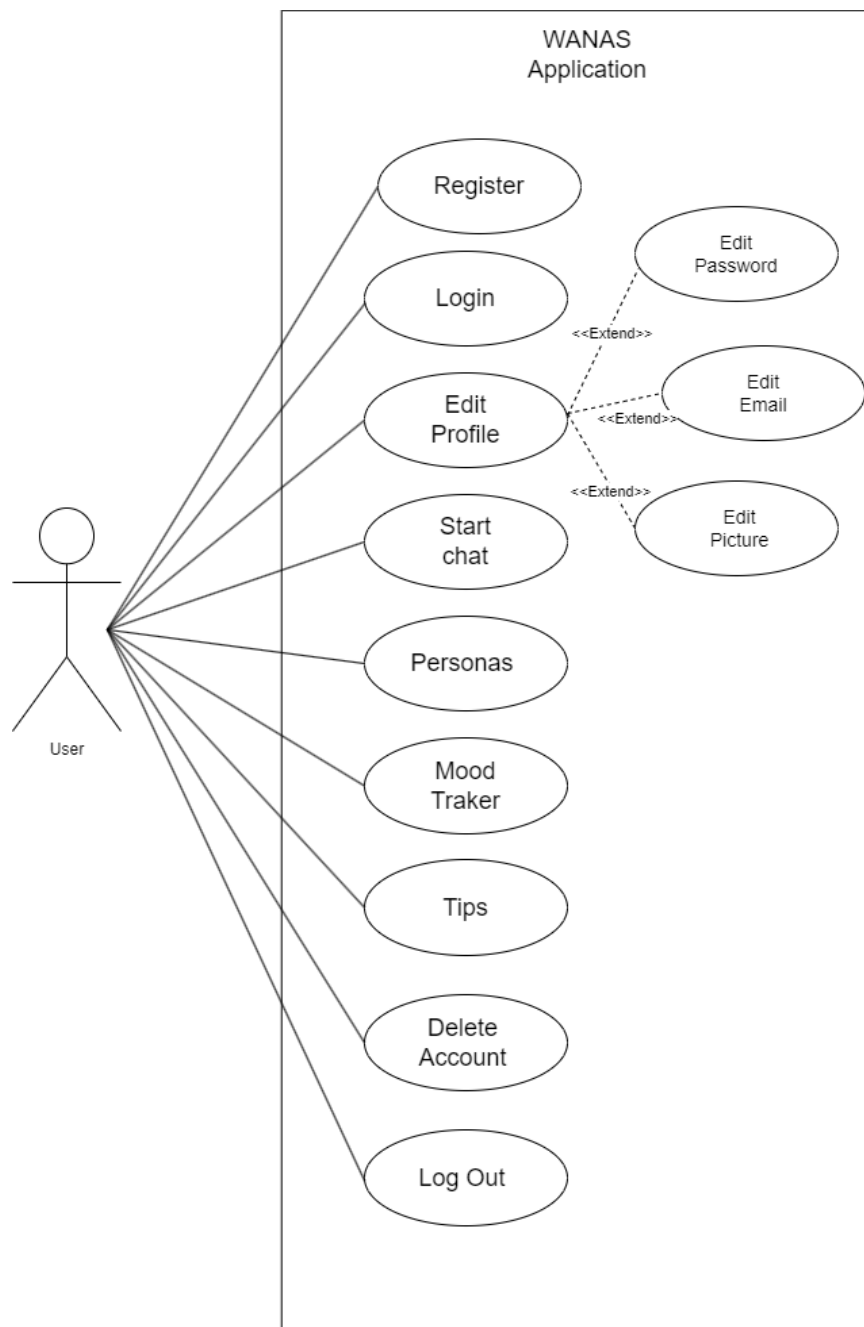
3.2   User Interface
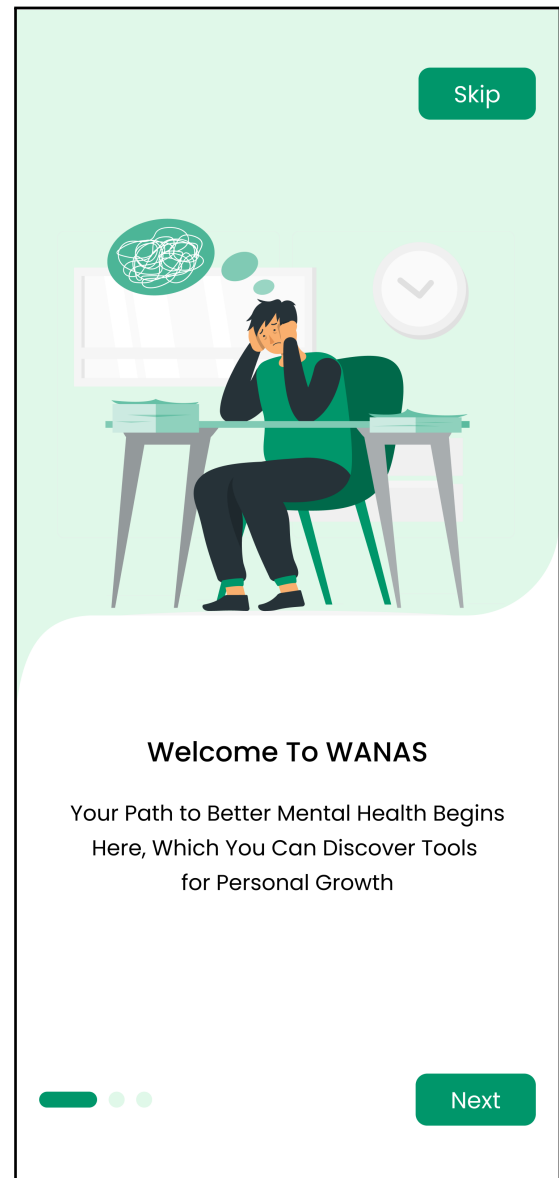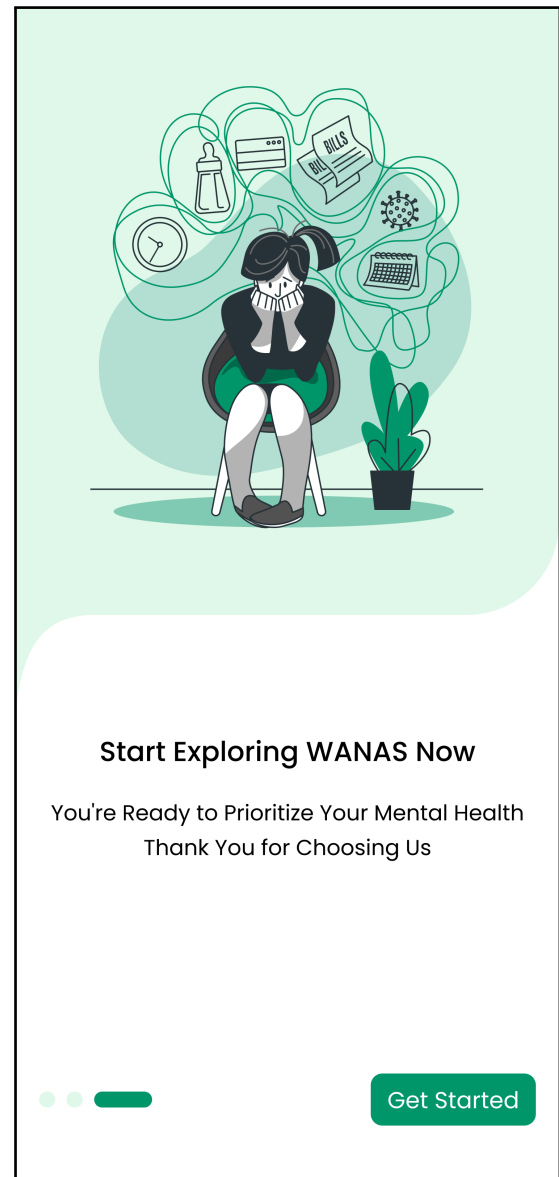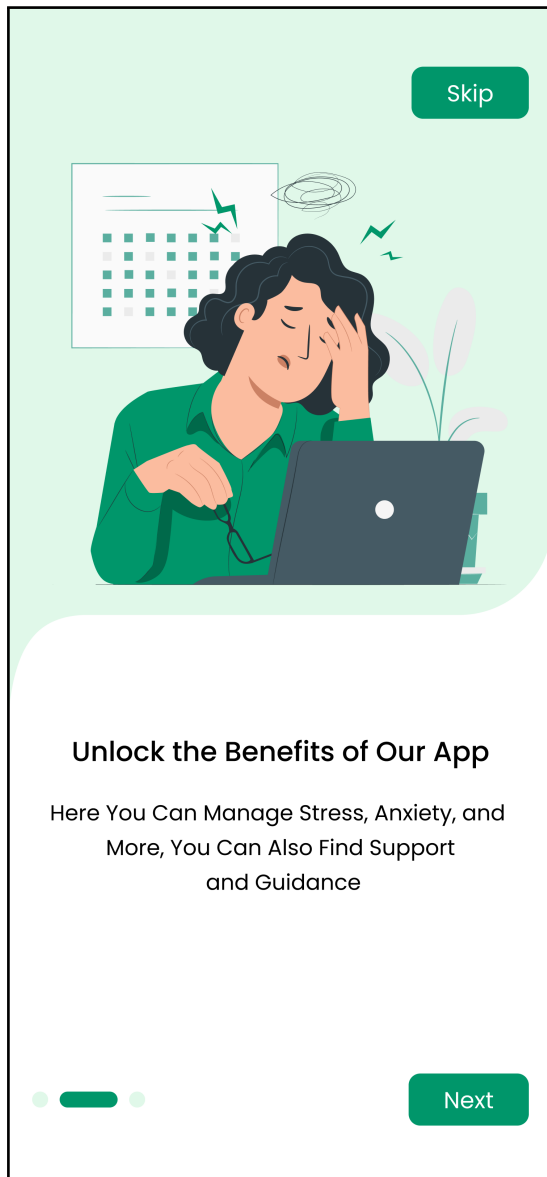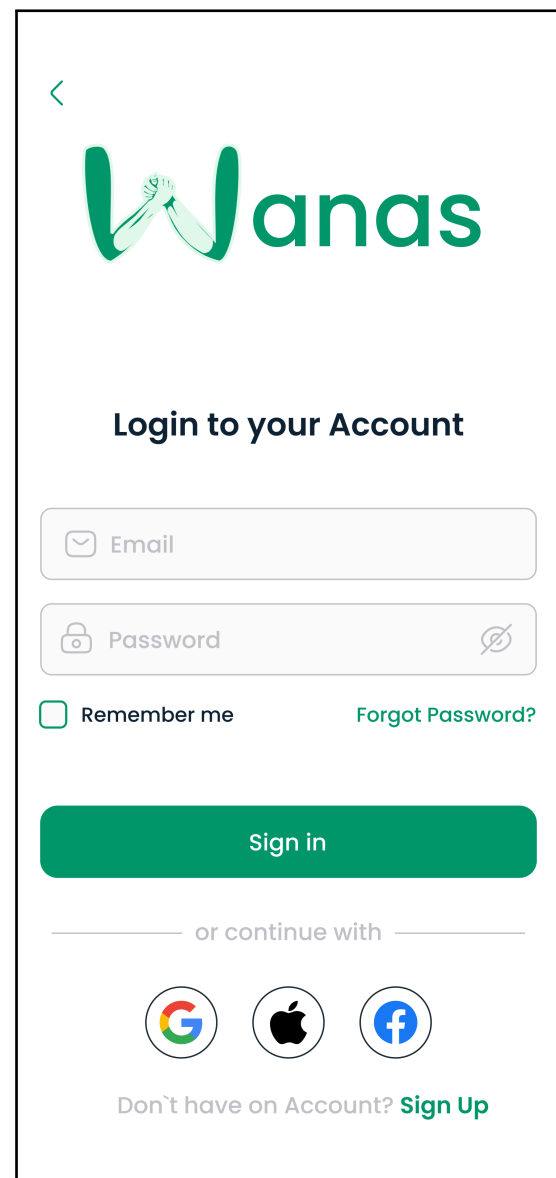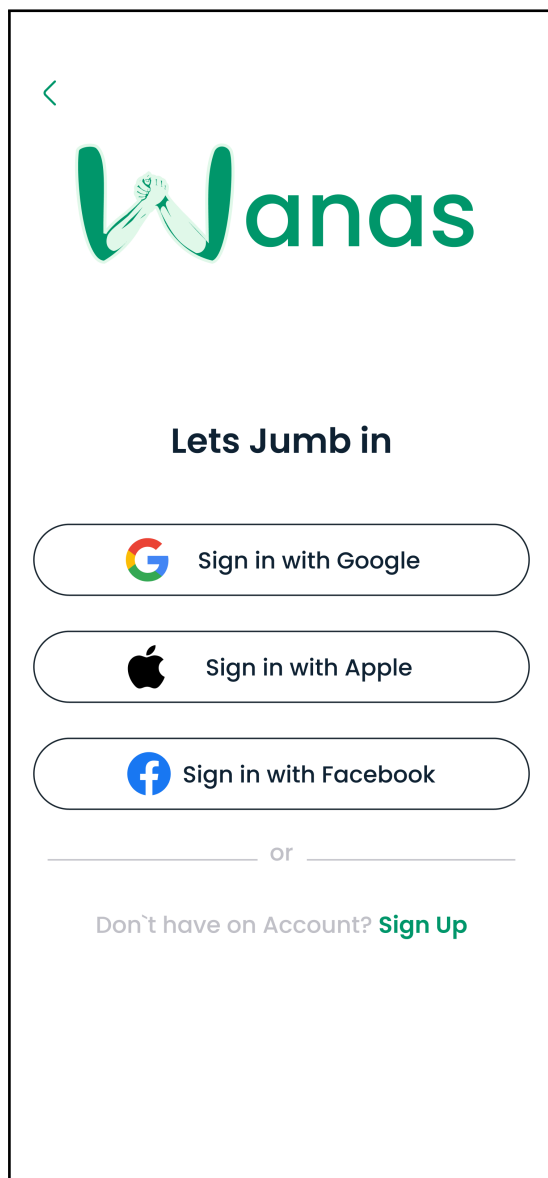
Unlock the Benefits of Our App

Here You Can Manage Stress, Anxiety, and
More, You Can Also Find Support
and Guidance

Skip

Next

Start Exploring WANAS Now

You're Ready to Prioritize Your Mental Health
Thank You for Choosing Us

Get Started
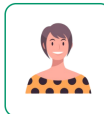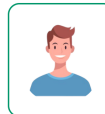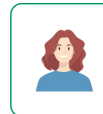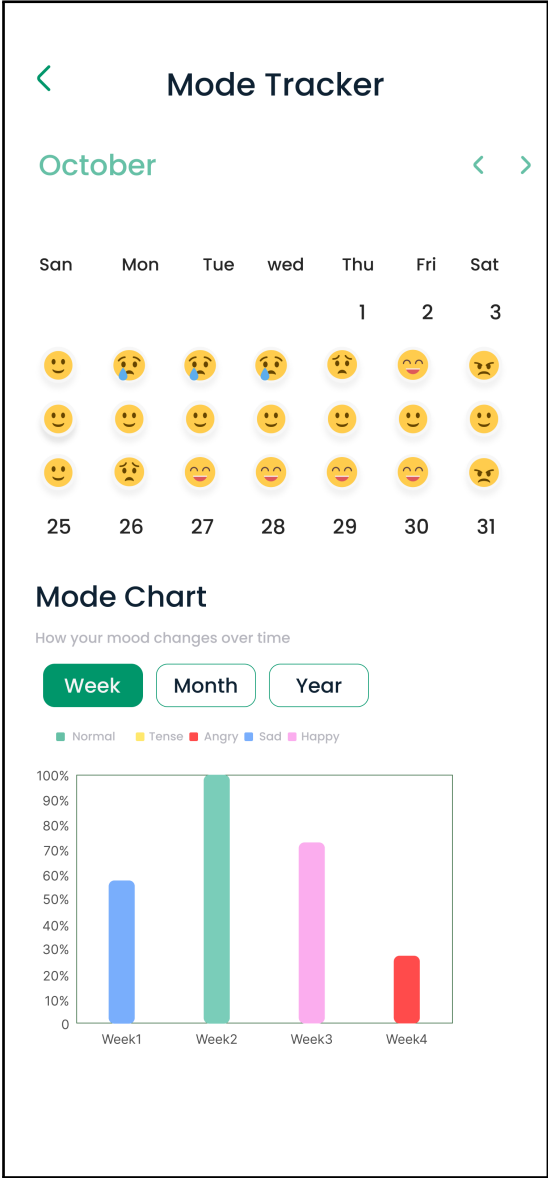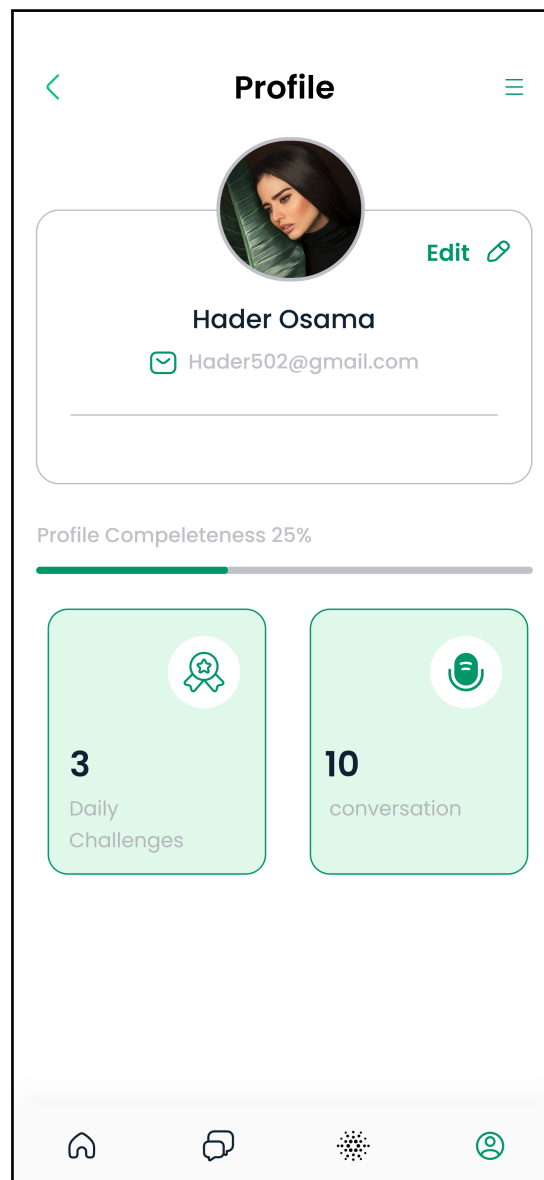
References

[1]   Shammur Absar Chowdhury et al. "Improving Arabic Text Categorization Using
        Transformer Training Diversification". In: *Proceedings of the Fifth Arabic Natural
        Language Processing Workshop*. 2020, pp. 226–236.

[2]   Diogo Cruz et al. "Reinforcement Learning Fine-tuning of Language Models is Biased
        Towards More Extractable Features". In: *Socially Responsible Language Modelling
        Research*. 2023. URL: https://openreview.net/forum?id=ee0kxTFS9a.

[3]   Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language
        Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter
        of the Association for Computational Linguistics: Human Language Technologies, Volume
        1 (Long and Short Papers)*. Ed. by Jill Burstein, Christy Doran, and Thamar Solorio.
        Minneapolis, Minnesota: Association for Computational Linguistics, June 2019,
        pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL:
        https://aclanthology.org/N19-1423.

[4]   Vasile Rus. "Natural Language Processing". In: *Encyclopedia of Sciences and Religions*.
        Ed. by Anne L. C. Runehov and Lluis Oviedo. Dordrecht: Springer Netherlands, 2013,
        pp. 1401–1404. ISBN: 978-1-4020-8265-8. DOI: 10.1007/978-1-4020-8265-8_1225. URL:
        https://doi.org/10.1007/978-1-4020-8265-8_1225.

[5]   Chi Sun et al. "How to Fine-Tune BERT for Text Classification?" In: *Chinese
        Computational Linguistics*. Ed. by Maosong Sun et al. Cham: Springer International
        Publishing, 2019, pp. 194–206. ISBN: 978-3-030-32381-3.

[6] Ibtissam Touahri. "AraBERT with GANs for High Performance Fine-Grained Dialect Classification". In: *Proceedings of the 6th International Conference on Big Data and Internet of Things*. Ed. by Mohamed Lazaar et al. Cham: Springer International Publishing, 2023, pp. 160–170. ISBN: 978-3-031-28387-1.

[7] Ashish Vaswani et al. "Attention is All you Need". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/ 3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.