UNIVERSIDADE
BEIRA INTERIOR

Conversational Agent for Mental Health Support

Luís Gonçalo Aguilar Silva

Dissertação para obtenção do Grau de Mestre em
Engenharia Informática
(2º ciclo de estudos)

Orientador: Prof. Doutor Sebastião Augusto Rodrigues Figueiredo Pais

Outubro de 2025

Declaração de Integridade

Eu, Luís Gonçalo Aguilar Silva, que abaixo assino, estudante com o número de inscrição M13747 do Mestrado em Engenharia Informática da Faculdade de Engenharia, declaro ter desenvolvido o presente trabalho e elaborado o presente texto em total consonância com o Código de Integridades da Universidade da Beira Interior.

Mais concretamente afirmo não ter incorrido em qualquer das variedades de Fraude Académica, e que aqui declaro conhecer, que em particular atendi à exigida referenciação de frases, extratos, imagens e outras formas de trabalho intelectual, e assumindo assim na íntegra as responsabilidades da autoria.

Universidade da Beira Interior, Covilhã 08/10/2025

Luis Gonçalo Aguilar Silva

Acknowledgements

This master’s degree journey was challenging, enlightening, and incredibly fulfilling. It put my resilience to the test, fuelled my curiosity, and served as a reminder that fulfilling work is rarely done alone. I am appreciative of everyone who supported me, gave advice when needed, and thought this project would have a tiny, useful impact.

Without the support and confidence of my supervisor, Professor Sebastião Pais, I could not have finished this dissertation or much of my academic career. His support, insightful criticism, and unwavering faith in the importance of this study influenced the dissertation and my development as a researcher. I am especially thankful for his availability during critical decision points.

I am deeply grateful to my family. To my parents and my brother: thank you for your unconditional support, patience, and the many quiet sacrifices that created space for me to learn.

To my friends and colleagues, thank you for the camaraderie, the honest feedback, and the good humour that made long days in the lab not only productive but also meaningful. Our conversations, were instrumental to the direction and quality of this work. I am equally indebted to my teachers across the years for instilling intellectual curiosity, methodological discipline.

I would also like to acknowledge the IT department and the university staff, administrative and technical, whose professionalism and kindness kept the practical side of research manageable. This work relied on the computing resources of the University of Beira Interior, and I am especially grateful to the UBI IT teams for providing hardware.

The machine-learning community’s openness was really helpful to this dissertation. Specifically, Hugging Face’s tools, datasets, and model hubs allowed for quick prototyping, benchmarking, and iteration. I express my gratitude to the authors of models published on the Hugging Face Hub, and the maintainers of open-source libraries, whose evaluation scripts, documentation, and checkpoints made reproducible experimentation and learning possible that would not have been possible otherwise.

Resumo

A saúde mental é hoje um dos maiores desafios globais, marcada pelo aumento de casos de ansiedade e depressão e pelas dificuldades de acesso a apoio especializado, custos elevados e estigma. Neste âmbito, as soluções digitais que expandem a abrangência dos cuidados constituem uma abordagem promissora. Esta dissertação apresenta a conceção e implementação de um agente conversacional para apoio à saúde mental. O modelo principal utilizado foi o Gemma 3 4B, aplicado com engenharia de *prompts*, a afinação Low-Rank Adaptation (LoRA) foi explorada mas não adotada devido a constrangimentos de exportação. Este Large Language Model (LLM) é complementado por modelos de *Hate Speech Detection*, *Emotion Detection*, *Sentiment Analysis* e *Sarcasm Detection*. Os classificadores BERT (sentimento, emoções, sarcasmo e ódio) geram sinais que condicionam o tom e a estratégia de diálogo, sendo essas respostas integradas no *system prompt* para que a interação seja dinâmica. A arquitetura suporta a entrada e saída de dados em texto e contém módulos para *check-ins* periódicos, deteção de sinais de risco de automutilação ou desesperança com oferta do contacto de linhas nacionais de apoio ao suicídio e gestão de medicação através do envio de prescrições em PDF e agendamento de lembretes para a toma de medicação. Experiências mostram que a metodologia proposta produz respostas mais empáticas, seguras e úteis do que usar apenas um modelo de linguagem em cenários variados. A similaridade semântica do coseno foi calculada para comparar o agente com aplicações do mercado e com cortes experimentais sem os módulos de análise, e as pontuações indicaram conteúdo comparável, enquanto as respostas do agente foram notavelmente mais validantes, empáticas, seguras e orientadas para a prática. Os componentes adicionados como a verificação regular com o utilizador e o agendamento de medicação para criar lembretes, indicam potencial prático como terapeuta virtual. As limitações incluem erros ocasionais de deteção de intenção se quer uma resposta direta ou para ajudar a refletir, perda de contexto devido à memória curta e componentes incompletos de fala e avatar.

Palavras-chave

Saúde Mental, Terapeuta Virtual, Chatbot, Inteligência Artificial, Agente conversacional incorporado

Resumo alargado

A saúde mental enfrenta um agravamento da carga de ansiedade e depressão, associado a barreiras de acesso, custos e estigma. Como resposta, foi concebido e avaliado um agente conversacional de apoio à saúde mental baseado no modelo Gemma 3 4B, orientado por princípios de segurança, ética e utilidade clínica. O sistema adapta o tom e a estratégia de resposta ao estado emocional do utilizador, realiza check-ins periódicos, reconhece sinais de risco e disponibiliza informação de contacto de linhas nacionais de apoio ao suicídio, facilitando ainda a adesão terapêutica através de lembretes de medicação.

A arquitetura opera em modo texto e organiza dois percursos principais: cold-start para criação do perfil e fase terapêutica com memória conversacional. Cada mensagem do utilizador é analisada por classificadores BERT de sentimento, emoções, sarcasmo e discurso de ódio, cujos sinais são incorporados em prompts de sistema especializados para síntese de notas, escolha do estilo de intervenção (reflexiva ou de resposta direta) e geração da resposta terapêutica. O sistema inclui módulos de utilidade clínica: check-ins automáticos após períodos de inatividade, deteção de risco com disponibilização de linhas nacionais de prevenção do suicídio e gestão de medicação através da extração de dados de prescrições em PDF e agendamento de lembretes.

A avaliação empírica comparou o agente com aplicações disponíveis no mercado e com variantes experimentais do próprio agente sem determinados módulos. Os resultados qualitativos indicaram respostas mais empáticas, validadoras, seguras e orientadas para a prática, com melhoria da utilidade percebida em cenários diversos. Para complementar, foi calculada a similaridade semântica do cosseno entre respostas, a qual apontou conteúdo globalmente comparável, em consonância com as observações qualitativas que favoreceram o agente completo.

Os achados sugerem potencial prático enquanto terapeuta virtual, particularmente pela combinação de apoio emocional consistente, gestão de risco e suporte à continuidade do cuidado, mantendo elevada disponibilidade. Foram, contudo, observadas limitações, nomeadamente erros pontuais na deteção da intenção do utilizador entre pedido de resposta direta e incentivo à reflexão, perda de continuidade temática em interações prolongadas devido à memória curta e ausência de componentes de voz e avatar.

Propõe-se, como trabalho futuro, o reforço dos mecanismos de deteção e adaptação, a expansão da memória conversacional para preservação de objetivos de longo prazo e a experimentação com modelos mais recentes.

Palavras-chave

Saúde Mental, Terapeuta Virtual, Chatbot, Inteligência Artificial, Agente conversacional incorporado

Abstract

Mental health is one of the greatest global challenges today, marked by an increase in cases of anxiety and depression and difficulties in accessing specialised support, high costs and stigma. In this context, digital solutions that expand the scope of care are a promising approach. This dissertation presents the design and implementation of a conversational agent for mental health support. The main model used was Gemma 3 4B, applied with prompt engineering. LoRA tuning was explored but not adopted due to export constraints. This LLM is complemented by models for Hate Speech Detection, Emotion Detection, Sentiment Detection, and Sarcasm Detection. The BERT classifiers (sentiment, emotions, sarcasm, and hate) generate signals that condition the tone and dialogue strategy, with these responses being integrated into the system prompt so that the interaction is dynamic. The architecture supports text data input and output and contains modules for periodic check-ins, detection of signs of risk of self-harm or despair with the provision of national suicide helpline contact details, and medication management through the sending of PDF prescriptions and scheduling of medication reminders. Experiments show that the proposed methodology produces more empathetic, confident, and useful responses than using only a language model in varied scenarios. The cosine semantic similarity was calculated to compare the agent with market applications and with experimental cuts without the analysis modules, and the scores indicated comparable content, while the agent's responses were notably more validating, empathetic, safe, and practice-oriented. Added components such as regular check-ins with the user and medication scheduling to provide reminders indicate practical potential as a virtual therapist. Limitations include occasional errors in detecting whether a direct response or reflection is desired, loss of context due to short memory, and incomplete speech and avatar components.

Keywords

Mental Health, Virtual Therapist, Chatbot, Artificial Intelligence, Embodied Conversational Agent

Contents

List of Figures

List of Tables

Acronyms List

AI	Artificial Intelligence
API	Application Programming Interface
CBT	Cognitive Behavioural Therapy
ECA	Embodied Conversational Agent
ICBT	Internet-based cognitive behavioural therapy
LLM	Large Language Model
LoRA	Low-Rank Adaptation
NLP	Natural Language Processing
UBI	University of Beira Interior
VR	Virtual Reality
VRET	Virtual Reality Exposure Therapy
WHO	World Health Organization

Chapter 1

Introduction

This document is the final project for the Master's degree dissertation in Computer Science at the University of Beira Interior. Titled "Conversational Agent for Mental Health Support," it focuses on the development of a chatbot in mental health contexts under the supervision of Professor Sebastião Pais.

Approximately one out of every eight people worldwide suffers from a mental disorder. Anxiety and depressive illnesses have been increasingly prevalent in recent years [1]. However, despite this growing demand for help, significant barriers to accessing professional mental health care persist, particularly due to a shortage of licensed psychologists [2] and the high costs associated with consultations. These challenges have led to a substantial increase in the need for alternative support options.

This dissertation addresses the aforementioned challenges by exploring the development of a conversational agent capable of providing responsive and supportive interactions to meet mental health needs. This chatbot has been designed with the objective of offering accessible, empathetic, and immediate assistance, thereby assisting individuals in managing feelings of depression, anxiety, and loneliness when professional help may be unavailable. By leveraging advancements in artificial intelligence (AI), this conversational agent aims to deliver preliminary support, promote self-care strategies, and encourage users to seek further help when needed, thus serving as a valuable resource in expanding access to mental health support.

1.1 Motivation

Mental health disorders have become one of the leading challenges of modern society, affecting millions of people across all ages and backgrounds. Conditions such as depression, anxiety, and loneliness can severely impact quality of life, relationships, and productivity. Despite the growing awareness of these issues, access to professional support remains limited. Shortages of trained specialists, high treatment costs, and social stigma continue to create barriers that leave many individuals without timely or adequate care.

In this context, technology offers a valuable opportunity to bridge part of the gap. Recent advances in AI and natural language processing (NLP) have enabled the creation of conversational systems that are capable of holding meaningful interactions and adapting to user needs. These systems cannot replace professional therapy, but they can provide accessible, immediate, and empathetic support that complements existing healthcare services. By offering structured guidance, encouraging self-care practices, and recognizing when additional help may be needed, conversational agents can become an important resource for those who might otherwise face their challenges alone.

The motivation for this dissertation is to explore how these technologies can be integrated into a system designed specifically for mental health support. By combining large language models (LLMs) with emotion detection, therapeutic strategies, and supportive functionalities, the goal is to create a conversational agent that helps reduce barriers to care, encourages proactive well-being, and expands access to assistance for individuals in need.

1.2 Objective

The objective of this dissertation is to create a chatbot that can converse via text and speech and then merge into a conversational agent with a physical embodiment. The primary purpose of this conversational agent is to serve as a virtual therapist, assisting people with their mental health by treating issues such as depression and anxiety. The chatbot will also be able to read medical prescriptions and immediately notify users when it is time to take their medication. The agent can respond via voice or text, ensuring accessibility and according to user preferences.

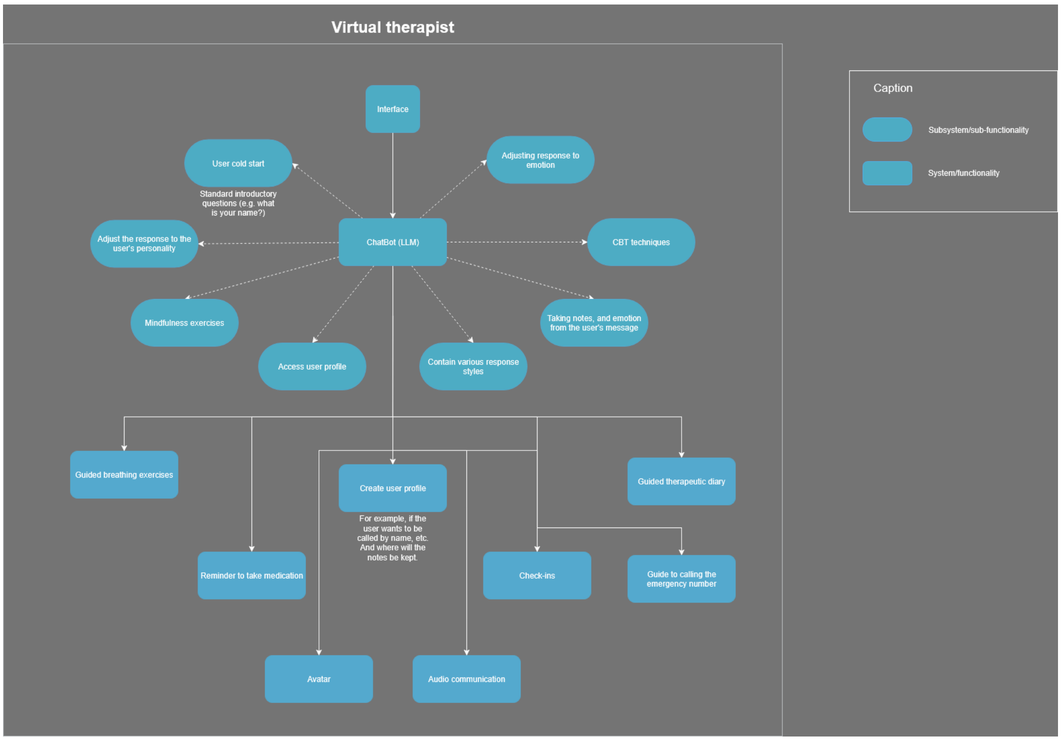


Figure 1.1: Diagram of the virtual therapist architecture

1.3 Dissertation Outline

1. **Introduction** - outlines the motivation behind the work, the research objectives, and the overall structure of the dissertation. It frames the importance of conversational agents in mental health contexts and defines the scope of the project.
    - Motivation

  - Objective
  - Dissertation Outline

2. **Background** - provides the necessary context by discussing the global mental health crisis, the concept of virtual therapists, and the role of Embodied Conversational Agents (ECAs). It also examines ethical considerations such as privacy, bias, and the limits of AI in mental health support.
    - Contextualization
    - Conclusion
3. **Related work** - surveys the literature on ECAs and large language models, with a particular focus on their applications in healthcare and mental health. It compares existing systems, discusses therapeutic chatbots, and highlights advances in sentiment, emotion, and text classification models relevant to conversational support.
    - Embodied Conversational Agents
    - Large Language Models
    - Embodied Conversational Agents in Health
    - Embodied Conversational Agents in Mental Health
    - Virtual Therapist
    - Conclusion
4. **Architecture** - presents the system's design and functionalities. It describes the modular architecture, including components such as the chatbot core, medication reminder, check-ins, suicide prevention module, avatar, audio interaction, guided diary, and breathing exercises.
    - Description of functionalities
    - Conclusion
5. **Methodology** - details the technical foundations of the implementation. It describes the tools used, criteria for selecting the models, the implementation of functional modules, and the prompting strategies. It also documents the main challenges encountered and the solutions adopted.
    - Tools used
    - Model Selection Criteria
    - Implementation of Functional Modules
    - Implementation Challenges and Solutions Adopted
    - Conclusion
6. **Experiments and Qualitative Discussion** - reports the evaluation of the system through different experimental cuts. It compares configurations (e.g., LLM-only vs. LLM with analysers) and discusses the impact of sentiment, emotion, hate detection, personalization, and response style on the quality of interactions.

  - Experimental Cuts
  - Discussion and Observation
  - Conclusion

7. **Conclusion** - summarises the main findings of the dissertation. It reflects on the extent to which the objectives were achieved, highlights the contributions made, and recognises the limitations of the current work.

8. **Future work** - proposes directions for extending the system, including building a user interface, training domain-specific models, expanding multilingual datasets, integrating vision-based prescription reading, and conducting clinical validation studies.

Chapter 2

Background

This chapter provides the necessary background to contextualise the research within the broader landscape of mental health and technological innovation. First, it examines the global and European burden of mental health disorders, focusing particularly on depression, anxiety and loneliness — conditions that significantly impair quality of life and place a strain on healthcare systems. Particular attention is given to the Portuguese context, where demand for mental healthcare exceeds available resources, thereby highlighting systemic barriers to timely and equitable access.

The chapter then explores how emerging technological interventions, such as virtual therapists and ECAs, can expand access to mental health support. From early experiments with systems such as ELIZA and PARRY, to contemporary applications that leverage virtual reality (VR) and internet-based cognitive behavioural therapy(ICBT), the evolution of conversational technologies demonstrates their increasing ability to provide meaningful psychological support.

Finally, the discussion turns to the ethical challenges of deploying these technologies, including concerns around safety, dependency, bias, privacy and equitable access. Together, these sections provide a foundation for understanding how chatbot interaction and prompt engineering can contribute to addressing mental health challenges while recognising the complexities and responsibilities involved in their implementation.

2.1 Contextualization

2.1.1 Mental health

Mental health is a growing public health concern in Europe, and mental health disorders place a considerable cost on healthcare systems, communities, and individuals. According to recent data, roughly 84 million people in the European Union suffer from mental health issues, accounting for nearly one-sixth of the region's population. Mood disorders, such as depression and anxiety, are among the most common, with depression accounting for a considerable proportion of disability in Europe. These disorders also hampers economic growth by reducing productivity, workforce involvement, and individuals' ability to fully contribute to the economy [3]. This highlights the urgent need for effective intervention and support mechanisms.

In Portugal, there has been an apparent growth in demand for mental health care. However, these programs are limited in both resources and accessibility. The low number of mental health specialists per population has resulted in long waiting lists, leaving many people without timely treatment. To illustrate, Portugal has around 2.5 psychologists per 100,000 people, which is well below the amount required to appropriately treat the increasing prevalence of mental health disorders[2].

2.1.1.1 Depression

Depression affects around 280 million individuals globally, according to the World Health Organization (WHO) [4]. It is a primary cause of disability worldwide and makes a considerable contribution to the overall disease burden. Depression affects people of all ages and socioeconomic backgrounds, impairing their capacity to function at job, school, and in personal relationships. The World Health Organization emphasizes that depression is more than just a passing sensation of melancholy or poor energy. It is a significant medical illness marked by continuous sadness, lack of interest in formerly enjoyable activities, changes in eating and sleep patterns, difficulties focusing, and even suicidal ideation. Severe cases might cause significant impairment in everyday living and functioning.

Further research conducted in 2022 revealed that approximately 42% of Portuguese adolescents reported experiencing depressive symptoms, representing a notable increase from previous years. This increase highlights the pressing necessity for the provision of mental health services that are specifically designed for adolescents, as early intervention and readily accessible support can help to mitigate the long-term effects of these challenges. It is imperative that this growing mental health crisis among young people be addressed in order to promote resilience and well-being as they transition into adulthood [5].

2.1.1.2 Anxiety

Anxiety disorders belong to the most frequent mental health diseases worldwide, impacting an estimated 301 million individuals in 2019. These disorders include a variety of conditions marked by excessive anxiety, concern, or behavioural difficulties that profoundly disrupt daily life. Anxiety disorders' symptoms frequently begin in childhood or adolescence and, if not treated, can last into adulthood, causing significant suffering and decreased functioning in personal, social, and professional domains. The WHO emphasizes that, while everyone has anxiety at times, people with anxiety disorders have chronic and uncontrollable fear or concern, which is frequently accompanied by physical symptoms such as tension, heart palpitations, nausea, or sleep difficulties. Anxiety disorders include generalized anxiety disorder, panic disorder, social anxiety disorder, phobias, and others, with many having numerous types at once. Despite the availability of very effective therapies such as cognitive-behavioural therapy and medication, approximately one in every four people suffering from anxiety disorders receives the care they require. Treatment barriers include a lack of awareness, insufficient mental health resources, stigma, and a paucity of educated healthcare practitioners[6].

Deco Proteste recently conducted a poll of 1,563 people in Portugal from September to December, finding that 57% of women reported mental health problems in the previous three months, compared to 35% of males. Anxiety was the most commonly reported problem, impacting 43% of women, twice as many as males. Young people aged 18 to 34 were most affected, with 70% of women and 47% of men reporting emotional problems. Despite this, just 43% of people who have experienced psychological problems in the last three years have sought help, with 56% attempting to address them on their own. Of those who sought help, 65% utilized medication, while 50% chose psychotherapy, which had the highest satisfaction rates. Consultations, while faster in the private sector, can cost an average of 137 euros every month[7].

2.1.1.3 Loneliness

According to a recent EU-wide poll undertaken by the Joint Research Centre (JRC), 13% of Europeans experience loneliness on a daily or near-daily basis. The findings emphasize that loneliness, which increased during the COVID-19 pandemic, is still a persistent problem [8]. Young adults aged 18 to 24 and senior adults over 65 are the most affected. Among young people, 20% report feeling lonely on a daily basis, which is much higher than in other age groups. In contrast, older adults, particularly those over the age of 65, frequently experience loneliness as a result of restricted mobility, retirement, or the loss of loved ones, with 15% experiencing it on a regular basis. Furthermore, single-person households and urban surroundings are more prone to experience loneliness, emphasizing the importance social connectivity[8].

The report underlines the link between loneliness and mental health. Individuals who frequently experience loneliness are up to three times more likely to report feelings of anxiety, despair, and low life satisfaction than those who do not[8].

2.1.2 Virtual therapist

A virtual therapist is a digital entity designed to deliver therapeutic interventions, often leveraging AI, VR, or other digital technologies to simulate or complement human therapy. Unlike telehealth therapists, who are licensed professionals conducting therapy remotely, virtual therapists may operate autonomously or semi-autonomously without direct human oversight[9][10].

For anxiety and phobia, virtual reality exposure therapy (VRET) has shown significant benefits. Studies report that VRET reduces phobic symptoms by providing controlled, immersive environments for gradual exposure, outperforming traditional methods in some cases[11][12].

* In depression management, ICBT supported by virtual therapists has achieved moderate to large effect sizes, particularly when therapist guidance is included. Blended approaches combining digital and face-to-face sessions have proven effective for patients unresponsive to medication[13][14].

Continued improvement phenomena, such as sustained symptom reduction post-treatment, have been observed in both anxiety and depression interventions[13][12]. These findings underscore the potential of virtual therapists to expand access to evidence-based care while maintaining clinical efficacy.

2.1.3 Embodied Conversational Agents

Joseph Weizenbaum built the first ”chatbot” for use in mental health environments, ELIZA, in 1966. It is widely considered as the first chatbot capable of participating in human-like conversation. Weizenbaum’s goal was not to create a therapeutic tool, but to investigate the possibilities of human-machine connection. ELIZA’s programming used a ”DOCTOR” script, which mimicked a Rogerian therapist by evaluating user input for keywords and providing responses that fostered open-ended conversation. This script intentionally modelled a non-directive therapeutic style, inviting users to elaborate on their thoughts and participate in self-reflection[15].

ELIZA provided two crucial insights regarding chatbot interaction and its usefulness in creating connections in mental health contexts. Initially, despite using simple pattern-matching algorithms, ELIZA was able to instil a deep sense of comprehension and connection in users, who were regularly compelled to share personal views. This answer highlighted the ability of even basic AI to create a sensation of being heard, suggesting that AI tools could improve mental health by supporting the experience of being acknowledged and affirmed [16].

In 1972, psychiatrist Kenneth Colby created PARRY, an early chatbot that aimed to emulate human communication by emulating the behaviour of a person with paranoid schizophrenia [17]. PARRY’s success in simulating psychiatric conditions with realistic interactions laid foundational ideas for the use of AI in mental health support, demonstrating how AI could offer empathetic and condition-specific responses to human users, bridging emotional gaps and aiding in understanding mental health [18].

2.1.4 Ethics

In designing and deploying a conversational agent aimed at mental health support, it is essential to affirm that this system is not a replacement for professional clinical care; it is a supportive tool intended to complement, but in no way substitute psychotherapy, psychiatric care, or any medical treatment. Users must always be made aware that if they exhibit signs of crisis, severe emotional distress, suicidal ideation, or any indications that exceed the agent’s scope, they should immediately seek help from qualified mental health professionals. Ensuring user autonomy and safety is fundamental, especially in such sensitive contexts.

From the outset, the agent must ensure that user data, particularly sensitive data concerning emotional state, psychological health or mental well-being, are processed in full compliance with applicable privacy standards. The EU’s General Data Protection Regulation (GDPR, Regulation (EU) 2016/679) applies across member states and mandates strong protections when handling personal data, especially “special categories” of data, which include health and mental health information. In Portugal, Law No. 58/2019 of 8 August ensures the implementation of the GDPR in national law and establishes the rules for legitimate processing, consent, security, confidentiality, and the rights of data subjects[19].

The ethical utilisation of ECAs and LLM within the context of mental health must be guided by fundamental principles such as beneficence, non-maleficence, autonomy, justice, privacy and transparency. Adhering to these principles ensures that these technologies prioritise user well-being, prevent harm, respect individual autonomy and promote equitable access to care[20][21].

Several key ethical challenges arise in this context. First, ensuring user safety is paramount, particularly in high-risk scenarios such as self-harm or severe psychological distress. Systems must incorporate robust protocols for crisis detection and escalation, including mechanisms that connect users to human professionals when necessary [21][22].

Second, there is a need to manage risks of dependency and ensure that these technologies complement rather than replace human care. Studies highlight the risk that users might over-rely on these tools due to their accessibility, anonymity, and cost-effectiveness, potentially neglecting the need for in-person therapy in complex cases requiring nuanced human judgment and emotional intelligence[23].

Third, bias mitigation remains a critical concern. Research indicates that LLMs can exhibit variations in empathy and responsiveness across demographic groups, raising questions about fairness and inclusivity [24][25]. To address this, systems must be trained using diverse datasets to prevent the reinforcement of stereotypes or the marginalisation of vulnerable populations[26][27].

Finally, given the sensitivity of mental health data, privacy and confidentiality are foundational requirements. Robust safeguards must be in place to prevent data misuse and ensure compliance with regulatory standards[21].

2.2 Conclusion

This chapter provides background information that highlights the scale of the mental health crisis and the opportunities presented by technological innovation. Mental health disorders such as depression, anxiety and loneliness are widespread and costly, yet they are not adequately addressed by existing healthcare systems, particularly in countries like Portugal where shortages of specialists limit access to timely care.

Against this backdrop, AI-powered interventions, including virtual therapists and emotional support assistants (ECAs), emerge as promising tools for providing additional support, enabling early intervention and offering scalable, cost-effective solutions. Their capacity to simulate therapeutic dialogue, deliver structured interventions, and provide constant accessibility underscores their potential to augment traditional care pathways.

However, integrating them into mental health support systems must consider ethical issues. Issues of safety, dependency, fairness and data protection must be rigorously addressed to ensure these technologies enhance rather than undermine user well-being. Situating chatbot interaction and prompt engineering within this broader context establishes the groundwork for subsequent discussions on the responsible design and deployment of these tools to alleviate the pressing challenges of modern mental healthcare.

Chapter 3

Related work

ECAs and recent advances in LLMs are converging to reshape how people access information, receive support, and learn in sensitive domains such as healthcare and mental health. ECAs bring multimodal, social interaction, speech, facial expression, gesture, and gaze, while LLMs provide flexible language understanding, reasoning, and generation. Together, they promise more capable, adaptive, and empathic systems; yet they also raise questions about safety, evaluation, cultural fit, and clinical validity. The purpose of this chapter is to synthesize prior work across these threads, to surface the assumptions and design choices that matter, and to identify gaps that motivate the research undertaken in this dissertation.

3.1 Embodied Conversational Agents

ECAs are animated, computer-generated characters designed to simulate human-like interactions through multimodal communication. By integrating speech, facial expressions, gestures, and eye gaze, ECAs create intuitive and engaging interfaces that bridge the gap between humans and machines [28]. Emerging from advancements in human-computer interaction, ECAs have evolved into sophisticated systems capable of fostering trust, empathy, and adaptability in diverse applications.

The design and implementation of ECAs rely on modular architectures that integrate multimodal interaction, fuzzy knowledge bases, and natural gesture synthesis to enhance user experience and functionality. These components work together to create lifelike agents capable of engaging in human-like communication.

Multimodal interaction methods enable ECAs to process and generate communication across multiple channels, such as speech, facial expressions, and gestures. This approach mirrors human conversational behaviour, making interactions more intuitive and engaging [28]. For example, synchronized verbal and non-verbal outputs, such as gestures and gaze, are essential for maintaining conversational flow and turn-taking [29].

3.1.1 Evaluation, Empathy, and User Interaction

Effective evaluation of ECAs requires a combination of methodologies that assess usability, trustworthiness, and user interaction. Eye tracking studies, for instance, reveal how users distribute attention during interactions, providing insights into engagement and perception. It has been demonstrated that users often treat ECAs as conversational partners, with gaze patterns aligning closely to those seen in human-human interactions[30]. Similarly, usability testing and self-report measures are effective in quantifying user satisfaction and task performance, as shown in previous work[31].

Empathic behaviours and subtle expressivity significantly enhance non-verbal communication and user perception. Yalçın and Leonhardt emphasize that subtle cues, such as variations in gaze direction, voice tone, and gesture speed, play a critical role in regulating social dynamics and fostering emotional connections. According to prior findings, these behaviours complement explicit expressions, contributing to a more natural and engaging interaction[32]. Furthermore, cultural nuances in emotional expressions underline the importance of tailoring ECAs to diverse user groups[33].

3.1.2 Applications and Case Studies

ECAs have demonstrated significant potential across diverse domains, including education, customer service, and privacy-sensitive technologies.

In language learning, ECAs have been employed as virtual tutors to improve vocabulary, pronunciation, and conversational skills through multimodal interactions. Systems such as Baldi and Bao utilize 3D animated talking heads, visual speech, and gestures to enhance user engagement and learning outcomes, showing improvements in speech articulation and linguistic awareness[34]. Similarly, ECAs integrated into collaborative systems have been found to enhance peer interaction and cognitive activity, contributing positively to learning outcomes [35]. In addition, intelligent tutoring systems like AutoTutor and Victor provide personalized feedback and emotional support, which foster deeper learning and increase user satisfaction[36][37]. However, it is worth noting that prolonged training periods are often required to achieve significant and lasting improvements [34].

Comparative studies highlight the advantages of multimodal ECAs over conventional interfaces. For example, ECAs like SIVA adapt their conversational style and expressive behaviours to match user preferences, resulting in higher perceived empathy and believability[38]. However, challenges remain in ensuring adaptability and addressing individual differences in user expectations[39]. These findings underscore the importance of tailoring ECAs to specific applications and user needs.

3.1.3 Large Language Models

Large language models demonstrate an impressive capacity for text-based generalization based on extensive data sets, rendering them highly versatile and efficient tools in a range of domains. In the field of education, they provide students with personalized assistance, developing study materials that are specifically designed to meet their individual requirements [40]. In the field of finance, they assist in the analysis of risks, the forecasting of market trends, and the automation of processes[41]. In the field of computer programming, LLMs facilitate the development process by automating tasks such as code generation, problem-solving, and optimization of complex operations[42].

Moreover, in the field of scientific research, these models have the potential to facilitate the acceleration of discoveries through the synthesis of articles, the interpretation of data, and the formulation of hypotheses. Another relevant application is decision automation and data analysis, where LLMs process large volumes of information, identify patterns, and offer actionable insights, allowing for faster and more informed decisions [43].

3.1.4 Large Language Models for Mental Health

LLMs in mental health have demonstrated promise in a variety of applications, including diagnosis, therapeutic intervention, and emotional support. These models have the ability to let users express themselves more effectively through conversational exchanges[44]. Emotional support systems like CASE has effectively modelled cognitive-affective interactions for empathic assistance [45], and SMILE has advanced multi-turn dialogue capabilities [46]. Despite these advancements, the datasets used in these solutions are frequently taken from online forums, which lack input from professionally accredited psychological practitioners [47].

ChatCounselor is a groundbreaking development. It uses an open-source LLM calibrated with the Psych8k dataset, which contains real counselling transcripts from certified professionals. This allows the model to deeply understand psychological principles. It also uses advanced data processing to turn important interactions into useful query-answer pairs, ensuring high-quality, relevant responses. This study also significantly contributed by creating specialized benchmarks, including the Counselling Bench. When tested across seven parameters like active listening, empathy, and information provision, ChatCounselor outperformed other open-source models like Alpaca-7B and Vicuna-v1.3-7B. These results highlight how crucial domain-specific fine-tuning and evaluation are for making LLMs more applicable in mental health [48].

Other initiatives also highlight the potential of LLMs in mental health. For instance, CBT-LLM fine-tuned a Chinese LLM with a Cognitive Behavioural Therapy (CBT) QA dataset to generate professional, structured responses aligned with CBT intervention strategies. Empirical evaluations showed CBT-LLM excelled in producing relevant and high-quality therapeutic responses, proving its practical use in mental health support [49]. Similarly, the ”Cactus” dataset introduced multi-turn dialogues based on CBT techniques, helping models like Camel surpass others in counselling skills by systematically applying CBT methodologies [50].

Furthermore, methods like the Chain of Empathy (CoE) prompting integrate psychotherapy models such as CBT to boost empathetic reasoning in LLMs. This approach leads to balanced and emotionally appropriate responses, emphasizing the importance of emotional context in AI-driven therapy [51].

3.1.5 Prompt engineering

Prompt engineering is a relatively emerging field that focuses on creating and optimizing prompts to efficiently use language models in a variety of applications and research areas. This field seeks to improve understanding of the capabilities and limitations of LLMs. Researchers use quick engineering to boost LLM performance in a variety of tasks, including question answering and arithmetic reasoning. Developers provide strong and effective prompting strategies for integrating LLMs with other technologies. Beyond creating prompts, prompt engineering involves a wide range of skills and techniques required for engaging with and expanding on LLMs. It is critical in assuring the safety of LLMs and establishing new capabilities, such as enhancing LLMs with domain knowledge and external tools[52].

Prompt engineering not only improves response quality and uniformity, but it also addresses flexibility and coherence, guaranteeing that LLM systems fulfil stringent precision and sensitivity requirements. By fine-tuning the triggers that guide chatbot responses, developers may create virtual assistants that reply appropriately and provide genuine emotional support, establishing a sense of understanding and security among users. These characteristics are particularly important in fields where ethical issues are vital, such as mental health. Furthermore, prompt engineering improves the adaptability of LLM across applications by eliminating uncertainty and enhancing dependability. According to Amazon Web Services, using sophisticated prompt techniques such as ”few-shot” or ”zero-shot” or ”Chain-Of-Thought” learning allows models to be quickly adapted to new jobs without the need for large training data. Prompt engineering not only refines model outputs to meet ethical norms, but also improves the model’s adaptability and responsiveness across various applications [53].

3.1.5.1 History

NLP has seen an enormous shift in the area of prompt engineering, moving from relatively simple language inputs to highly sophisticated, context-sensitive models. The first developments in this field can be traced back to the introduction of early language models and retrieval systems that used statistical and keyword-based methods, laying the groundwork for subsequent advances in prompt engineering that have come to shape the behaviour of modern AI models[54].

Before the introduction of neural networks, basic language models such as n-grams and early information retrieval systems relied on user-defined keyword prompts to retrieve data from text collections. As neural networks gained prominence between 2010 and 2015, important advancements such as ”Word2Vec” and Sequence-to-Sequence models aided the creation of contextual word embeddings and machine translation, laying the groundwork for more advanced prompt engineering[54].

A significant change occurred in 2015 with the introduction of attention mechanisms. Transformers changed how prompts might guide model outputs by prioritizing pertinent text input segments. This allowed for improved context interpretation. By allowing model outputs to be adjusted depending on reward functions, reinforcement learning has advanced the field by 2017. This improved quick controllability and addressed issues like exposure bias and model bias[54].

Significant progress was made by Prompt Engineering in 2018 with the introduction of BERT, which used bidirectional contextual encoding to support transfer learning and task-specific fine-tuning. Control codes and template-based prompting were introduced the next year, allowing prompt engineers to provide more focused outputs and improve interpretability and model adaptability[54].

LLMs with significant parameter counts, as GPT-3, were introduced in 2020 and 2021. In order to reduce bias, prompt engineering techniques expanded to include a range of styles, domain-specific modifications, and morally sound tactics. Recent developments have improved model answers through context-aware and intent-based tactics, such as multimodal prompting, multi-turn conversation handling, and better personalization features [54].

3.2 Embodied Conversational Agents in Health

In healthcare, ECAs are increasingly used to support patients in managing chronic conditions, providing disease-specific knowledge, and facilitating interactive learning [55]. Their human-like qualities can improve trust and satisfaction, especially when designed with empathetic behaviours and secure information handling[56]. For example, ECAs can deliver personalized health advice or monitor patient progress in real-time, leveraging their multimodal capabilities to adapt to individual needs[57]. Their design and implementation continue to evolve, emphasizing trustworthiness, ease of use, and emotional connection[58].

3.2.1 Applications of Embodied Conversational Agents in Healthcare

For elderly patients, ECAs serve as social companions, alleviating loneliness and promoting cognitive health. For example, Daisy, an ECA designed as a potted flower, engages older adults in casual conversations and suggests activities to enhance social connectivity [59]. Additionally, ECAs are being explored as virtual speech therapy assistants for individuals with neuro degenerative disorders, offering personalized exercises and monitoring progress[60]. In periconception and pregnancy care, ECAs provide tailored guidance and emotional support. A study at Erasmus MC demonstrated their potential to improve trustworthiness and patient satisfaction through empathetic interaction styles[58].

3.2.2 Psychological and Social Implications

The psychological and social impact of ECAs in healthcare is significantly influenced by their visual and behavioural attributes, which shape user perceptions of personality, trust, and social interaction. ECAs are often perceived as conversational partners, with users responding to them similarly to human interactions due to their multimodal communication capabilities, including speech, gestures, and facial expressions[30]. However, trust levels can vary based on the agent's appearance and communication style, as demonstrated in healthcare settings like periconception care[58].

Personality traits of ECAs, shaped by their design and behaviour, also influence user engagement. Emotional models incorporating personality traits improve social presence and interaction quality[61].

3.2.3 How do Embodied Conversational Agents compare to traditional therapy methods?

Empirical evidence suggests that ECAs can achieve comparable therapeutic outcomes to traditional psychological therapies in certain contexts, but their effectiveness varies depending on the intervention type and user engagement. Controlled trials and systematic reviews indicate that ECAs are effective in reducing psychological distress, particularly in structured interventions targeting mood and anxiety disorders[62]. However, studies comparing ECAs to active control groups, such as traditional therapy, often fail to demonstrate superior effects[62].

User engagement and satisfaction are critical factors influencing the success of ECAs. Reviews highlight that ECAs are generally well-received, with positive usability and acceptability outcomes[63]. Their ability to simulate human-like interactions through verbal and non-verbal cues enhances user comfort and adherence, addressing common barriers in self-guided interventions[64]. However, ECAs may struggle to replicate the depth of emotional connection achieved in traditional therapy settings[65].

Some of the advantages of ECAs are their accessibility, as they can provide mental health support to people who might not have access to traditional therapy due to location, financial constraints, or social barriers. For example, ECAs effectively reduced stress among university students during the COVID-19 pandemic by offering culturally tailored and personalized support[66]. Additionally, anonymity is a key benefit, as users often feel more comfortable sharing sensitive information with ECAs, which can reduce stigma and increase engagement[67].

3.2.4 Existing Studies, Applications, and Implementations

In recent years, there has been a growing interest in the use of ECAs and chatbot technologies to support various mental health needs. These systems have been explored across a range of contexts, from anxiety and depression to suicide prevention and general well-being. Table 3.1 summarizes a selection of notable studies and applications that demonstrate the diversity of implementations and therapeutic goals in this emerging field.

Name of Study or App	**Mental Health Context**	**Description**
Woebot	Anxiety and Depression	Woebot is an AI-powered chatbot that delivers CBT techniques to alleviate symptoms of anxiety and depression. It provides structured conversations and evidence-based tools for mental health support [68].
Assessment of users' acceptability of a mobile-based ECA for the prevention and detection of suicidal behaviour	Suicide prevention, crisis intervention	This paper presents HelPath, a mobile app using an ECA to monitor suicidality risks and support users with CBT-based guidance, showing promising results in a pilot study on user acceptance and adherence [69].
Replika	Empathetic responses, Loneliness	Replika is a social chatbot designed to provide emotional support and companionship. It helps users cope with loneliness and mental well-being through empathetic interactions [70].
ECAs Providing Motivational Interviewing to Improve Health-Related behaviours: Scoping Review	General mental well-being	Promotes health-related behaviours; motivational interviewing for reducing alcohol use [71].

Table 3.1: Summary of studies and applications using conversational agents in mental health contexts

3.3 Virtual therapist in Mental Health

Virtual therapists are computer-based systems or virtual agents designed to provide therapeutic support, guidance, or treatment to individuals. These systems are often used in various fields, including mental health, physical rehabilitation, and speech therapy. They aim to simulate the role of a human therapist by offering interaction, feedback, and personalized care, often through VR, AI, or other digital platforms.

3.3.1 Possible functionalities of a virtual therapist

Adjust the Response to the User Emotion: Virtual therapists use sentiment analysis and emotional intelligence to detect user emotions, such as sadness, anxiety, or frustration, through text, voice, or facial expressions. By analysing linguistic cues and interaction patterns, the system adjusts its tone and content to provide empathetic and contextually appropriate responses. For example, during moments of distress, the chatbot may adopt a soothing tone and suggest calming exercises [72].
Cold Start of User: To address the challenge of limited initial user data, virtual therapists employ cold start strategies. These include asking introductory questions to understand the user's mental health needs, preferences, and goals. The system may also use general therapeutic techniques, such as mindfulness or psychoeducation, until sufficient data is gathered to personalize interactions[73].
CBT Techniques: Virtual therapists implement CBT by guiding users through structured exercises such as cognitive restructuring, journaling, and behavioural activation. These techniques help users identify and challenge negative thought patterns, promoting healthier coping mechanisms. The chatbot provides real-time feedback and adapts exercises based on user progress[73].
Be Able to Give Various Styles of Therapy: To cater to diverse user needs, virtual therapists offer multiple therapeutic approaches, including CBT, mindfulness-based therapy, and solution-focused therapy. Users can select their preferred style, or the system can recommend one based on the user's profile and emotional state. This flexibility ensures a tailored therapeutic experience[74].
Taking Notes of Users' Messages: Virtual therapists maintain a record of user interactions to track progress and provide continuity in therapy. These notes include key themes, emotional states, and recurring issues, enabling the system to offer personalized follow-ups and insights. This feature also supports long-term progress tracking[75].
Access to User Profile: The system uses a user profile to store information such as therapy goals, preferences, and interaction history. This profile allows the virtual therapist to personalize responses, recommend exercises, and adapt therapy plans dynamically. Access to this data ensures a consistent and user-centred therapeutic experience[72].

3.3.2 Therapeutic Chatbots

Therapeutic chatbots are AI-powered conversational agents designed to provide mental health support and other therapeutic interventions. They use of NLP to simulate human-like conversations and offer assistance in areas such as anxiety, depression, stress management, and behavioural changes. These chatbots often incorporate evidence-based techniques like CBT and mindfulness exercises to deliver personalized support and coping strategies [76]. The concept of therapeutic chatbots began with ELIZA in 1966, a program simulating a Rogerian psychotherapist. ELIZA demonstrated the potential for computers to engage in therapeutic dialogue but lacked true understanding or therapeutic efficacy[77].

3.3.3 Text Classification Architectures for Conversational Systems

Emotion and sentiment analysis in conversational systems has evolved significantly, transitioning from traditional lexicon-based methods to advanced deep learning architectures. Early approaches relied heavily on lexicon-based techniques, which used predefined sentiment dictionaries to assign polarity scores to text. While these methods were interpretable and domain-independent, they struggled with scalability, context sensitivity, and handling linguistic nuances like sarcasm or ambiguity[78][79]. Hybrid models that combined lexicons with machine learning partially addressed these limitations but still fell short in accuracy and adaptability[80][81].

Classical machine learning models, such as Support Vector Machines and Logistic Regression, introduced supervised learning to sentiment analysis, leveraging labelled datasets for improved performance. However, these models required extensive feature engineering and were limited in capturing sequential dependencies in text[82][83].

The advent of recurrent neural networks and their variants, such as Long Short-Term Memory networks, marked a turning point. These architectures excelled in modelling sequential data, enabling more nuanced emotion recognition in conversational contexts[84][85]. Despite their success, challenges like vanishing gradients and computational inefficiency persisted, paving the way for transformer-based models like BERT, which have since set new benchmarks in emotion and sentiment analysis[86][87].

Transformer-based models, such as BERT and RoBERTa, excel in sentiment analysis due to their self-attention mechanisms, bidirectional context understanding, and robust pretraining objectives. These features enable them to outperform traditional and recurrent architectures in capturing linguistic nuances and long-range dependencies in text[86][88].

The performance gap is particularly pronounced in multilingual and domain-specific tasks. For example, in Urdu Named Entity Recognition, XLM-RoBERTa achieved an F1-score of 0.9969, outperforming hybrid recurrent models[89].

For this dissertation, BERT-based models were used, given their strong performance in capturing contextual nuances and improving emotion and sentiment analysis in conversational systems.

3.4 Conclusion

The literature points to three converging trajectories. First, ECAs have matured from animated front-ends to interactional partners whose credibility depends on tight coupling of verbal and non-verbal behaviour, fine-grained timing, and context-sensitive expressivity.

Second, LLMs have dramatically expanded the linguistic competence and adaptability available to such agents, provided that their outputs are shaped through careful prompting, tooling, and safety scaffolds. Third, in health and mental health, early deployments show encouraging gains in engagement and short-term outcomes, but also reveal persistent gaps in personalization, cultural attunement, transparency, and clinically grounded evaluation.
Across studies, several limitations recur. Evidence is often short-term, with small or convenience samples, and heterogeneous outcome measures that complicate comparison. Many systems rely on forum-derived data or synthetic dialogues rather than clinician-authenticated interactions, limiting clinical validity. Personalization is frequently shallow (e.g., surface-level style changes) rather than principled adaptation to user goals, history, and affect. Safety practices (risk detection, escalation, and auditability) are uneven, and privacy expectations are not consistently operationalized. Finally, evaluation rarely integrates behavioural signals (e.g., gaze, prosody), user-reported measures (e.g., alliance, trust), and task outcomes into a unified framework.

Chapter 4

Architecture

The architecture of the proposed system forms the structural and functional basis for the development of a virtual therapist, designed to provide personalised and adaptive psychological support. This system integrates multiple interdependent modules, which together enable the analysis of user data, the interpretation of emotional states, the selection of appropriate therapeutic strategies, and the delivery of empathetic responses in real time.

The architectural design was guided by two fundamental principles: personalization and modularity. Personalization ensures that each interaction is tailored to the user's profile and emotional state, taking into account their history, preferences and the current situation. Modularity, meanwhile, guarantees that the various components ranging from the NLP engine based on LLM to integrated sub-features such as adaptive emotion response generation, cold start user integration, medication reminders and guided diary logging can be refined or expanded without affecting the overall system's functionality. This approach enables the architecture to remain flexible and scalable, keeping pace with technological and scientific advances while maintaining consistency in the therapeutic experience.

This architecture was also designed to support different modes of interaction, namely text, voice and visual elements, allowing for an immersive and inclusive experience. In addition, the presence of mechanisms for recording and continuously monitoring interactions ensures therapeutic consistency and enables longitudinal monitoring of the user's progress.

As illustrated in the figure 1.1, the configuration of the components and the flow of information between them is depicted, with emphasis placed on the interactions between the ChatBot's central module LLM, the personalization and emotional analysis subsystems, the therapeutic intervention modules, and the user communication interface.

4.1 Description of functionalities

4.1.1 Interface

The User Interface acts as the entry and exit point of the system. It supports both text-based and voice-based communication. Incoming user inputs are captured and forwarded to the processing pipeline, while system responses are rendered back to the user either as text or synthesized speech. In configurations that include an avatar, the interface also handles the synchronization of visual outputs such as facial expressions and gestures.

4.1.2 ChatBot

At the core of the architecture lies the ChatBot component, which is powered by a LLM. This module is responsible for understanding and generating natural language, receiving preprocessed input from the user interface and producing contextually appropriate responses. It operates under system prompts that encode therapeutic strategies and behavioural guidelines, ensuring alignment with the intended support objectives. The LLM also orchestrates sub-functions such as mindfulness routines, check-ins, therapeutic exercises, generating textual content that other modules then transform into interactive experiences.

4.1.2.1 Emotion Adaptation

This module uses sentiment and emotion analysis techniques on user input. By identifying emotional states such as anxiety or sadness, it provides the LLM with signals that guide the response generation process. The module's purpose is not to replace the LLM, but rather to provide additional contextual variables that enable the system to adjust its tone, style and content.

4.1.2.2 Cold start

The onboarding process is triggered when a new user initiates a session. This module collects initial data, such as the user's name, communication preferences, and goals. These parameters are stored in the user profile and serve as the foundation for subsequent personalization. The module ensures that even in the absence of historical data, the system can provide relevant and structured support.

4.1.2.3 User profile

The user profile is a persistent data structure that stores contextual information across sessions. This information includes conversation history, detected emotional trends, preferences and, if applicable, medication schedules. The profile is regularly updated and accessed by other modules to ensure consistent responses and continuity across multiple interactions.

4.1.2.4 Mindfulness exercises

The Mindfulness Exercises sub-functionality provides guided practices that are integrated into the chatbot's workflow. It generates structured prompts to guide users through predefined mindfulness routines, thereby supporting stress reduction, improved focus and better emotional regulation. The module can be triggered either by a user request or by contextual conditions detected by the LLM. Once invoked, it provides sequential instructions through the conversational interface, which can be combined with audio or visual output to enhance interaction.

4.1.3 Medication Reminder

The Medication Reminder module manages the user's medication from start to finish, from capturing and interpreting prescriptions to creating reminders and confirming intake. The aim is to improve therapeutic adherence by providing timely notifications and a structured record of doses, which can be used by the rest of the system for personalisation and longitudinal monitoring. This functionality is included in the proposed system's scope and includes reading prescriptions and sending alerts at the appropriate time.

4.1.4 Check-in

The Check-in module monitors user interaction patterns and manages periods of inactivity. It uses activity logs to identify periods of inactivity and, if no interaction is registered for two consecutive days, automatically triggers a notification inviting the user to provide an update. The module communicates with the LLM to generate the check-in prompt content and ensure its delivery through the interface. Operating as a background service enables it to maintain proactive engagement and guarantee regular user contact.

4.1.5 Suicide Prevention Hotlines

The Suicide Prevention Hotlines module recognises signs of self-harm or suicidal thoughts in user interactions. Whenever such signs are detected, the system immediately provides the user with the contact details of the appropriate suicide prevention hotline for their country or region. While the conversational agent can offer empathetic support, it cannot replace professional clinical care. Therefore, this module ensures that users immediately receive reliable information on where to seek specialised help at critical moments.

4.1.6 Avatar

The Avatar module generates a visual, multimodal representation of the conversational agent. It synchronises facial expressions, body movements and lip articulation with the dialogue stream in real time to complement textual and auditory outputs. The module integrates with the LLM by receiving generated responses and speech synthesis, mapping them to animation signals to drive visual behaviours. This ensures that the language output is temporally aligned with the avatar's behaviour, thereby enhancing user engagement and supporting a coherent conversational flow.

4.1.7 Audio communication

The Audio Communication module facilitates two-way interaction via speech. User input is captured and transcribed into text using Automatic Speech Recognition (ASR) technology, before being processed by the a LLM. System responses are then converted into audio output using text-to-speech, enabling seamless vocal communication. The module integrates directly with the dialogue pipeline to ensure real-time synchronisation between spoken and textual interaction streams.

4.1.8 Guided Therapeutic Diary

The Guided Therapeutic Diary module offers structured journaling functionality integrated into the chatbot. It guides users through predefined prompts and exercises to encourage self-reflection and emotional awareness. The LLM dynamically generates adaptive questions based on the conversational context and presents them via the interface. User entries are stored as structured data, enabling longitudinal tracking and personalised dialogue in subsequent interactions.

4.1.9 Guided Breathing Exercises

The Guided Breathing Exercises module provides interactive breathing routines, which can be initiated either by the user or via contextual cues. The LLM provides sequential instructions through text, audio or visual channels. This component controls the pacing and structure of the exercises to provide step-by-step guidance and synchronisation with the multimodal outputs. It operates as a lightweight coping mechanism, effortlessly integrated into the conversational flow.

4.2 Conclusion

The architecture provides a flexible framework for the virtual therapist, ensuring clear separation of components such as the LLM core, emotion adaptation, personalisation and multimodal interaction. This modularity allows subsystems to be refined independently, while persistent user profiling ensures continuity and adaptive personalisation. Critical safety mechanisms, including suicide prevention and medication management, are tightly integrated to meet reliability and ethical standards. Overall, the design provides a robust basis for implementation and future scalability, as well as the seamless integration of emerging therapeutic tools and technologies.

Chapter 5

Methodology

This chapter details how the proposed system was built and deployed, translating the architectural goals into concrete, testable components. It begins by outlining the software stack and supporting libraries selected to ensure modularity, scalability, and dependable runtime behaviour (Section 5.1, Table 5.1). It then motivates the model choices that underpin language understanding and text analytics, including the primary LLM as well as specialised classifiers for sentiment, emotions, sarcasm, and hate speech (Section 5.2).

Subsequently, the chapter describes the implementation of the functional modules that realise the end-to-end workflow. These include dataset curation and normalisation for instruction-style learning (Section 5.3.1); parameter-efficient fine-tuning with LoRA, together with the associated hardware, training configuration, and outcomes (Section 5.3.2); and the FastAPI service that orchestrates the conversational pipeline, state management, background scheduling, and document processing (Sections 5.3.3–5.3.4). Particular attention is given to the layered prompting strategy used to control behaviour across phases cold start, note-taking, intervention selection, therapeutic response, and medication extraction (Section 5.3.5). Finally, the chapter reports the main engineering challenges model export and deployment, stability under shared hardware, and context-window limits and the pragmatic solutions adopted to keep the system reliable in practice (Section 5.4).

Overall, the chapter provides a comprehensive account of the design choices and trade-offs that enabled a safe, multilingual, and resource-aware therapeutic assistant, connecting tools, models, and APIs into a cohesive implementation.

5.1 Tools used

The implementation of the system necessitated the integration of many tools, programming languages, and libraries to ensure modularity, scalability, and consistent performance. Visual Studio Code, a versatile environment that facilitated coding, debugging, and testing, was used for development.

The main framework for running and managing LLM was chosen to be Ollama. To facilitate seamless communication between Ollama-running models and back-end services, the Python Ollama client library was incorporated. Furthermore, models based on the Transformer architecture were supported by the Transformers library (Hugging Face), enhancing the system's capacity for NLP.

The FastAPI framework was used to define and expose RESTful endpoints in the Python implementation of the back end. Uvicorn served as the ASGI server, guaranteeing the effective operation of the application programming interface (API).

The APScheduler library (BackgroundScheduler) was integrated with Python's built-in datetime and timedelta modules to schedule time-sensitive or recurring activities, like medicine reminders or inactivity check-ins.

The extraction of structured text from PDF documents was made possible by the io library's support for PDF upload and parsing in conjunction with pdfplumber.

The tools and libraries utilized in the implementation are summarized in the table 5.1.

Tool / Library	**Purpose**
Visual Studio Code	Main IDE for development, debugging, and testing
Python	Core programming language for the back-end and module integration
FastAPI	Framework for building the RESTful API
Uvicorn	ASGI server used to run the FastAPI application
Ollama	Framework to run LLMs locally
Ollama Python Library	Python integration for interacting with Ollama models
Transformers (Hugging Face)	Execution of models based on the Transformer architecture
APScheduler (Background-Scheduler)	Scheduling notifications such as medication reminders and check-ins
datetime, timedelta	Native Python libraries for handling time and date operations
io	Utility library for file input/output operations
pdfplumber	Reading and extracting structured text from PDF documents

Table 5.1: Summary of tools and libraries used in the implementation

5.2 Model Selection Criteria

5.2.1 Large Language Model

The first stage of the project's development involved selecting the most suitable LLM. After evaluating various open-source options, Gemma 3 4B was chosen over alternatives such as LLaMA 3.2 3B. One of the main reasons for this decision was Gemma 3's ability to handle an extended context window of up to 128,000 tokens, as this feature is crucial for tasks that require large-scale context integration. Furthermore, Gemma 3 introduces multimodal capabilities, including vision input, making it more versatile than purely text-based models. Gemma 3 is also one of the most recent models available, having been officially released by Google in March 2025, whereas LLaMA 3.2 was released in September 2024.

Another important consideration was the size of the model. With approximately one billion more parameters than LLaMA 3.2 3B, Gemma 3 4B provides greater representational capacity while remaining within the memory constraints of the available hardware. From a practical perspective, Gemma 3 4B proved to be a resource-efficient model, running effectively on a machine with a single GPU, which is an important factor given the limitations of the available hardware.

During practical testing, Gemma 3 4B also demonstrated better adherence to system prompts in certain text-processing scenarios than LLaMA 3.2, contributing to improved consistency and alignment with project requirements. Benchmark results further reinforced this decision: Gemma 3 4B achieved superior performance in critical reasoning and mathematical tasks.

Its recent architecture, extended context capacity, additional parameters, improved prompt-following behaviour, multimodal support, efficient deployment and strong benchmark performance all make Gemma 3 4B the most appropriate model for this project.

The decision to employ models based on BERT rather than GEMM3 was primarily influenced by the superior processing speed of the former for classification tasks. This preference is particularly salient in the context of the system's requirement for fast responses. The following essay will provide a comprehensive overview of the relevant literature on the subject.

5.2.2 Sentiment Model

In order to perform sentiment analysis, a suitable model was required that could provide both accurate results and multilingual support. As a significant number of sentiment analysis models available on Hugging Face provide insufficient documentation on their model card pages, the selection was primarily guided by quantitative indicators. These indicators included the number of downloads, benchmark results, and the range of supported capabilities. Following a thorough evaluation of the available criteria, the model `tabularisai/multilingual-sentiment-analysis`[1] was selected on the basis of its proven performance, widespread adoption within the relevant community, and its capacity for multilingual interpretation. This decision was made with the objective of identifying a model that would meet the specific requirements of the project.

The model under discussion provides a classification into five distinct sentiment categories: The following categories are employed: 'Very Negative', 'Negative', 'Neutral', 'Positive', and 'Very Positive'. This five-class configuration was regarded as being of particular benefit since it serves to not only ascertain whether a sentiment is positive or negative, but also to ascertain the intensity of the sentiment, whilst incorporating a 'Neutral' class for cases where the text does not express a clear polarity.

A significant additional factor in this selection was the model's multilingual capability. Given the multilingual nature of the project, incorporating texts in various languages, the utilisation of a multilingual sentiment analysis model is essential. This ensures broader applicability and more robust generalisation across diverse linguistic inputs. Consequently, this model was selected as the most appropriate choice to meet the project's requirements in terms of accuracy, granularity, and language coverage.

5.2.3 Emotion Model

In order to capture the emotional dimension of text, it was necessary to implement an emotion classification model. In order to complete this task, the model `AnasAlokla/multilingual_go_emotions_V1.2`[2] was selected and implemented through the Hugging Face pipeline pipeline("text-classification") functionality.

The selected model demonstrates the capacity to detect 27 discrete emotions in conjunction with a 'Neutral' category. This decision was made subsequent to a thorough consideration of

---

[1] `https://huggingface.co/tabularisai/multilingual-sentiment-analysis`

[2] `https://huggingface.co/AnasAlokla/multilingual_go_emotions_V1.2`

the available alternatives, which included options with either 6 or 27 categories. The model incorporating a greater number of classes was favoured on the basis that, although certain emotions may be semantically related or fall within the same broad category, they nevertheless represent discrete affective states. A finer-grained classification system facilitates more precise identification of subtle emotional differences, thereby enabling the LLM to tailor its responses with greater accuracy.

Another reason for this decision pertains to the scale and quality of the training data. The model underwent a process of fine-tuning on the *GoEmotions* dataset, originally developed by Google, utilised for emotion detection. Furthermore, the dataset was substantially augmented using LLM, resulting in a very large training set: approximately 6,619,306 samples for training, 825,643 for testing, and 825,643 for validation. This extensive dataset imbues the model with strong generalisation capabilities and robustness across diverse inputs. It is noteworthy that the reported accuracy for all emotion classes exceeds 0.98, further underscoring the reliability and efficacy of the model.

From a technical perspective, the model was fine-tuned from the base architecture, which is referred to as `google-bert/bert-base-multilingual-cased`. This approach is essential for the project, as the data may include text from different languages.

To summarise, the selection of this model was guided by several factors. Firstly, a fine-grained 27-class taxonomy of emotions was made available. Secondly, the robustness of the training on an extensively augmented dataset was taken into consideration. Thirdly, the exceptionally high accuracy across all emotion classes was a determining factor. Fourthly, the multilingual capability was a key consideration, and fifthly, the foundation on a reliable BERT architecture was deemed to be a significant advantage. Collectively, these characteristics render it a robust option for accurately capturing the subtleties of human emotion, thereby enabling the overall system to generate more contextually aligned and emotionally intelligent responses.

5.2.4 Sarcasm Model

Sarcasm detection was incorporated into the system to enhance interpretation of text where the intended meaning differs from the literal wording. To this end, the helinivan/multilingual-sarcasm-detector[3] model was selected and implemented via the Hugging Face pipeline(”text-classification”) functionality.

This model was chosen primarily because it is one of the most widely adopted sarcasm detection models on Hugging Face, with the highest number of downloads in its category. This reflects strong community trust and reliability. It is also multilingual, offering direct support for English, Dutch and Italian. While its performance is strongest in these three languages, fine-tuning from the base model `bert-base-multilingual-uncased` enables partial support for other languages, albeit with reduced accuracy.

The model was fine-tuned using the widely used English-language *News Headlines Dataset for Sarcasm Detection*, which provides high-quality data for learning the nuances of sarcastic expression. The fact that this dataset focuses on sarcasm in natural, real-world text makes

---

[3] `https://huggingface.co/helinivan/multilingual-sarcasm-detector`

it particularly well-suited to improving downstream tasks in sentiment or emotion analysis, where sarcasm can otherwise lead to misclassification.

In terms of performance, the model achieves impressive evaluation results: an F1 score of 87.23%, precision of 88.65%, recall of 86.33%, and overall accuracy of 88.30%. These metrics highlight a robust balance between precision and recall, confirming the model's reliability for sarcasm detection across multiple languages.

In summary, the model was selected due to its popularity, adoption and multilingual capability, as well as its foundation on a multilingual BERT architecture and strong quantitative performance metrics.

5.2.5 Hate Model

The model for hate speech detection selected and implemented was `christinacdl/XLM_RoBERTa-Multilingual-Hate-Speech-Detection-New`[4]. It used the Hugging Face pipeline `pipeline("text-classification")` functionality.

The selection was made in accordance with the same criteria applied in previous subsections: multilingual capability, benchmark performance, and the number of downloads. These factors reflect both technical suitability and community adoption. The model categorises text into two distinct classifications: The concepts of *Hate* and *No-Hate* are examined here. Despite its seemingly rudimentary binary structure, this approach has been demonstrated to be highly effective for downstream applications where the primary requirement is the identification of toxic or hateful content and the differentiation of this from acceptable text.

This model is based on the XLM-RoBERTa architecture, which is recognised for its strong performance across a variety of natural language understanding tasks and provides robust multilingual support. Consequently, it can be reliably applied in contexts where input text may be drawn from different languages.

In terms of performance, the model achieves competitive results, with a loss of 0.5873, a micro F1-score of 0.9065, a macro F1-score of 0.9050 and an overall accuracy of 0.9065. These high evaluation scores indicate that the model offers consistent detection across both hate and non-hate classes while maintaining a strong balance between precision and recall.

5.3 Implementation of Functional Modules

5.3.1 Dataset Preparation

The initial phase of the implementation process involved the collection and preparation of the datasets necessary for the training and evaluation of the models. A combination of publicly available datasets from both Kaggle and Hugging Face was utilised, with the focus on mental health conversations, counselling dialogues, and psychology-related texts. It is important to note that all of the datasets were originally in English and together amounted to a total of approximately 49,999 instances.

---

[4] `https://huggingface.co/christinacdl/XLM_RoBERTa-Multilingual-Hate-Speech-Detection-New`

Each instance consisted of two key components: a *prompt*, which described the problem or the user's condition, and a *completion*, which represented responses obtained either from chat-based interactions or from professionals in the field. This configuration enabled the data to be utilised for instruction-style fine-tuning and evaluation of language models.

In order to ensure consistency, the datasets were standardized by aligning their column structures into two unified fields: The first component, labelled 'input', is mapped to prompts or questions, while the second component, *output*, is mapped to answers or responses. The data cleaning steps that were performed included the removal of duplicates, the elimination of incomplete entries, and the harmonisation of inconsistent column names across different sources. Following the execution of these preliminary processing operations, the datasets were consolidated into a unified, integrated corpus.

The final step in the process involved converting the unified corpus into JSON format. In this format, each entry explicitly defined a `prompt` and a corresponding `completion`. The format was designed to streamline subsequent model training and inference, whilst also ensuring interoperability across the various frameworks and libraries utilised within the project.

The datasets utilised for corpus creation can be viewed in Table 5.2, which provides a synopsis of the diverse sources collected from Hugging Face and Kaggle, accompanied by their respective descriptions.

Dataset	**Source**	**Description / Content**
EmoCareAI/Psych8k	Hugging Face	Psychology-related dialogues for mental health support
ShenLab/MentalChat16K	Hugging Face	Multi-turn conversations focused on mental health topics
samhog/psychology-10k	Hugging Face	Psychology and counselling-related Q&A pairs
jkhedri/psychology-dataset-split	Hugging Face	Curated psychology dataset split for training and evaluation
Amod/mental_health_counseling_conversations	Hugging Face	Counselling-style mental health conversations
Riyazmk/mentalhealth	Hugging Face	Short dialogues related to mental health topics
mpingale/mental-health-chat-dataset	Hugging Face	Chat-based mental health conversation data
KelvinTichana2/mentalhealthcurated	Hugging Face	Curated mental health conversations
Kiran2004/MentalHealthConversations	Hugging Face	Counseling conversations focusing on mental health conditions
MaggiePai/mental_health_counseling_conversations	Hugging Face	Professional counselling-style conversations
Estwld/empathetic_dialogues_llm	Hugging Face	Empathy-focused dialogues to train emotionally aware responses
melissamonfared/mental-health-counseling-conversations-k	Kaggle	Counselling and therapy-related dialogues
zuhairhasanshaik/datacsv	Kaggle	General conversational dataset with mental health focus
thedevastator/synthetic-therapy-conversations-dataset	Kaggle	Synthetic dataset simulating therapy conversations
rahulsundkar/mental-health-conversational-dataset	Kaggle	Mental health dialogues designed for conversational AI

Table 5.2: Summary of datasets collected from Hugging Face and Kaggle for corpus creation

5.3.2 Model Fine-Tuning

The second stage of development involved fine-tuning the chosen LLM, `google/gemma-3-4b-it`, to align it with the project's domain-specific requirements. This was achieved using LoRA adapters, which were applied to specific layers of the model rather than updating all parameters. This approach was chosen for efficiency reasons: fully fine-tuning a 4B parameter model would be extremely resource-intensive in terms of GPU memory, training time and storage requirements. In contrast, LoRA drastically reduces both compute and memory costs while achieving comparable performance.

Hardware setup

The fine-tuning process was performed on a high-performance machine equipped with an NVIDIA RTX A6000 GPU with 50 GB of VRAM, 1 TB of system RAM and dual Intel Xeon Gold 5515+ CPUs.

Training setup

The model was fine-tuned using the Hugging Face `Transformers` and `TRL` libraries in combination with the `PEFT` framework for parameter-efficient training. Training was carried out with the following configuration:

- **LoRA configuration:** rank $r = 32$, $\alpha = 16$, dropout = 0.1, targeting the q_proj and v_proj modules of the transformer layers.
- **Training arguments:** batch size = 1 per device (train and eval), gradient accumulation steps = 4, optimizer = `paged_adamw_32bit`, learning rate = $1e^{-4}$, warmup steps = 10.
- **Training duration:** 30 epochs (total training time: approximately 7 days).
- **Evaluation metric:** `eval_loss` minimisation as the criterion for best checkpoint.

Results

The fine-tuning process produced a customised model checkpoint named `Gemma-3-4B-QA-therapeutics\_lora\_adapters-5-e5`. Training logs confirmed stable convergence across epochs. At the end of 30 epochs, the model achieved:

- **Evaluation loss:** 1.38
- **Mean token accuracy:** 0.699 (approximately 70%)

These results show that the Gemma 3 4B, adapted using LoRA, successfully specialised for the task at hand while maintaining stable performance and avoiding overfitting.

In summary, adopting LoRA adapters enabled the efficient fine-tuning of a 4B parameter model within the available hardware budget, achieving high performance while avoiding the prohibitive cost of full parameter updates. This parameter-efficient strategy enabled domain-specific knowledge to be integrated into the model while retaining scalability and ensuring practical deployment feasibility.

Limitations and Deployment Issues

Although the fine-tuning process was completed successfully, the resulting model was not used in the final system. The main reason for this was that the model needed to be converted to the `GGUF` format to be integrated with Ollama. However, this conversion could not be performed reliably due to technical constraints and a lack of access to the necessary export tools on the training server.

Additional challenges were also encountered during the fine-tuning stage itself. On several occasions, the training server became unstable and crashed, causing interruptions. Furthermore, as the hardware resources were shared, memory availability was sometimes limited, which prevented the model from being loaded or trained continuously. These factors, combined with the difficulties in model export, ultimately prevented the deployment of the fine-tuned version of Gemma 3 4B in the implementation.

5.3.3 API

The service runs on an ASGI server (`uvicorn`) and is implemented using `FastAPI`. It checks the most recent chat to initiate follow-ups if necessary and sets up background schedulers for notifications during launch.

Overview and Flow

The API orchestrates two conversational phases:

1. **Coldstart** — When no user profile exists or the profile is marked as incomplete, the request is routed to the coldstart routine to elicit essential profile information and initialize persistent memory.
2. **Therapy** — Once the profile is complete, subsequent messages are handled by the therapy routine, which uses conversational memory for continuity.

Endpoints

POST `/analyze_and_llm` This is the primary conversational entry point. The endpoint requires a JSON object with a single field, `text`, containing the user's input message. The response is also returned as a JSON object with one field, `response`, which holds the assistant's reply. If the user profile is missing or incomplete, the request is directed to the cold-start routine; otherwise, it proceeds through the therapy routine.

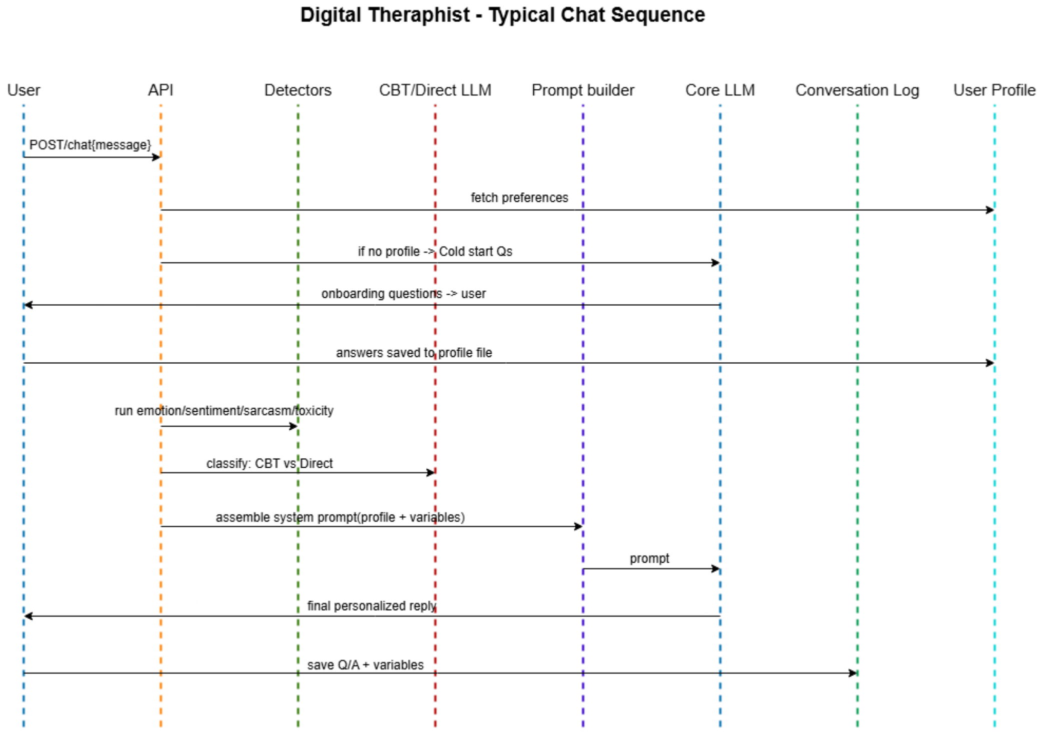


Figure 5.1: Sequence diagram of a typical chat interaction

When a message arrives, the API first checks the user profile's state. If no profile exists or the cold start routine has not been completed, the message is routed to the cold start function. Otherwise, the request is handled by the `therapy` function, which organises the full conversation workflow.

The therapy pipeline begins with a layer of *text analysis*. Each message is enriched by running it through auxiliary classifiers that detect sentiment, emotion, sarcasm, and hate speech. These signals provide affective and pragmatic context, allowing the LLM to adjust its response not only to the literal meaning of the text but also to the user's emotional state and tone. This analytic stage ensures that the system is not "blind" to subtler cues in the dialogue.

After analysis, the system constructs responses using a series of system prompts. In this phase, Ollama is called **three times per interaction**:

1. first with a **note-taking prompt** to summarise the user's input into concise notes,
2. then with an **intervention-mode prompt** to determine the appropriate style of reply (e.g., CBT-oriented, or direct),
3. and finally with a **therapy prompt** that integrates profile data, the outputs of the analysers, the intervention choice, and conversational history to generate the final response.

Thus, three distinct prompts structure the model's behaviour, each guiding the system at a different stage of processing.

Maintaining context across sessions is another key aspect of this pipeline. A rolling memory of up to thirty exchanges is managed by the system. After this limit is reached, old messages are deleted, but the memory is saved to disk so that discussions can continue without interruption even if the API is restarted. The *profile* that was generated during coldstart is constantly referred to in addition to memory; information such the user's name, preferred method of address, medical conditions, prescription drugs, and objectives are automatically entered into prompts. This guarantees that responses remain consistent with earlier information and personalised to the individual.

To ensure operational stability, two monitoring mechanisms are included. First, token accounting estimates the number of tokens consumed in each interaction, keeping usage within the model's context window. Second, all interactions are recorded in a structured log containing the timestamp, user input, analyser outputs, generated notes, and final response. These logs are essential both for transparency and for later evaluation of therapeutic performance.

In summary, the `therapy` function is not a simple relay but an orchestration layer that combines:

- classifier outputs for affective context,
- three successive system prompts with distinct roles,
- personalised information from the user profile,
- rolling conversational memory, and
- monitoring through token tracking and structured logging.

This layered design ensures that each user message leads to a coherent, emotionally aware, and personalised response, fully aligned with the project's therapeutic objectives.

POST `/upload_prescription` This endpoint accepts a medical prescription in PDF format. The file is processed with `pdfplumber` to extract text, which is then interpreted to identify prescribed medications. The response is returned in JSON format and contains a confirmation message along with the extracted medication details. Only PDF files are supported.
To enable the system to automatically read the drug regimen and set reminders, users can upload a prescription in PDF format. In order for the assistant to convert unstructured prescriptions into actionable messages, the design aimed to streamline the patient workflow.
The endpoint first verifies the uploaded file upon receiving a request. Currently, only PDF files can be uploaded; any other type will display an obvious error message. The `pdfplumber` library is then used to parse the PDF's content page by page, extracting the text for further examination.
The extracted text is passed to the function `interpret_medication_data`, which uses the Gemma 3 model through Ollama. A specialised *medication prompt* structures the input, instructing the LLM to output the result as JSON with the relevant details of each prescribed medication (e.g., name, dosage, timing, and any additional notes). The function also contains robust handling to extract JSON even if the model's output is wrapped in markdown blocks, ensuring that results are consistently machine-readable.
Once the JSON output is validated, it is merged with any existing data stored in `data/medication.json`. This ensures that new prescriptions can be appended to previously saved information rather than replacing it entirely. The updated file is then persisted to disk, providing a central record of all medications.
After updating the medication list, the system triggers the notification scheduler. Notifications are managed by APScheduler, which creates background jobs for each scheduled dose. For every medication entry:

- the number of daily doses is determined by the list of timings,
- each timing is parsed into hours and minutes,
- a cron-style job is registered to send a notification at the correct time,
- each notification includes the medication name, the dose number out of the total daily doses, and any additional notes provided in the prescription.

If invalid times are detected, they are safely ignored without interrupting the scheduling process.
In summary, the `/upload_prescription` endpoint automates the pipeline:

1. validating and extracting text from a prescription PDF,
2. interpreting the extracted data into a structured medication schema using the LLM,
3. merging the new data with the persistent medication file, and
4. scheduling notifications for each prescribed dose.

This integration of document parsing, language model reasoning, persistent storage, and background scheduling makes the endpoint a crucial feature of the system, bridging raw prescriptions and real-world medication adherence support.

State and Memory

The service persists a user profile in the data directory. Routing logic checks:

1. *Profile file exists?* If not, coldstart is invoked.
2. `coldstart_complete` *flag?* If `false`, coldstart continues; otherwise, messages go to therapy.

Conversational continuity is provided by `TherapyMemory` with a cap of 30 messages.

Implementation Notes

- **Startup hooks:** `@app.on_event("startup")` initializes medication schedules and triggers follow-up checks.
- **Validation:** request/response schemas are enforced via `pydantic` models (`TextInput`, `AnalysisResult_and_llm`).
- **PDF handling:** only PDF is accepted; text is extracted pagewise to reduce OCR variance and simplify parsing.
- **Error surfacing:** internal exceptions are propagated as HTTP errors with informative messages for observability.

5.3.4 Notifications

The system has an integrated notification mechanism to encourage involvement and adherence. The two functions of notifications are to remind users to take their prescription drugs and to inquire about their health following extended periods of inactivity. Both processes are background tasks that are initialized during API launch.

Medication reminders

Medication scheduling is triggered whenever a prescription is uploaded or when the service is restarted. The scheduler reads the `medication.json` file and creates jobs based on the timing fields associated with each medication. For every dose:

- the scheduler registers a cron-style job at the correct time of day,
- the notification includes the medication name, dose number, total doses for the day, and any additional instructions,
- invalid or malformed times are safely ignored without interrupting the process.

This ensures that users receive timely reminders aligned with their prescription, which is particularly important for adherence in therapeutic or medical contexts.

Inactivity check-ins

In addition to the generation of medication reminders, the system performs periodic well-being check-ins. Upon initiation, the function designated as `check_last_conversation` retrieves the stored conversation history and determines the timestamp of the most recent exchange. In the event that more than two days have passed since the previous interaction, a check-in message is triggered to encourage the user to re-engage (e.g., *”It’s been 2 days since your last conversation. May I inquire as to your well-being? Do you require a discussion?”).* In the event that the previous interaction was more recent, the system does not generate a notification, despite the fact that it does record this information.

This check-in mechanism is designed to ensure that periods of prolonged inactivity are identified and that a safety net is provided for users who may disengage during periods of difficulty.

Medication JSON schema

The interpreted prescription data is stored in a JSON structure. Each entry follows the schema shown in Table 5.3.

Field	**Type**	**Description**
`name`	String	Name of the medication (e.g., “Paracetamol”)
`dosage`	String	Dosage instructions, typically quantity and strength (e.g., “500 mg”)
`frequency`	String	frequency instructions, how many times to take a day (e.g., “2 times a day”)
`timing`	Array of strings	List of administration times in 24h format (e.g., [“08:00“, “20:00“])
`details`	String (optional)	Additional instructions such as “after meals” or “with water”

Table 5.3: Schema of the JSON object used for storing medications

5.3.5 System prompts

A crucial design choice in the implementation was the use of structured adaptive *system prompts*. Rather than relying on a single generic instruction, the system applies several specialised prompts that guide the behaviour of the language model depending on the conversational phase and task. Each prompt encapsulates a set of rules or goals, ensuring that the model’s outputs remain consistent, safe, and aligned with therapeutic objectives.

Therapy prompt

The `therapy_prompt` is the most comprehensive prompt in the system. It integrates multiple inputs—including user profile details (such as name, treatment preference, conditions, medication, and country), as well as outputs from the sentiment, emotion, sarcasm, and hate detectors—into a single instruction.

This prompt establishes the model’s role as an empathetic and safe mental health support assistant. The platform enforces a stringent set of guidelines, including the stipulation that the assistant does not supersede the role of professional assistance, the validation of emotional responses, the adaptation of communication styles to user preferences, and the management of sensitive cases such as self-harm or suicidal ideation by providing crisis helpline numbers appropriate to the user's country of residence.

In addition, it adapts its tone based on the selected *intervention mode*. If "CBT" is chosen, the assistant uses open-ended, reflective questioning. If "direct" is chosen, it provides concrete suggestions and practical strategies. Regardless of mode, mindfulness exercises are always offered if the user reports difficulty concentrating, memory issues, or requests relaxation techniques.

Coldstart prompt

The `coldstart_prompt` is used only during the initial profiling phase. Its sole purpose is to ensure that the user's responses to profiling questions (e.g., name, preferred form of address, health conditions, medication, goals, and country) are converted into a valid JSON object.

The prompt explicitly instructs the model to always respond in JSON format and provides clear schemas for each question. If the user's answer is unclear or irrelevant, the prompt directs the model to return an empty JSON or an error message, always preserving machine-readability. This structured approach guarantees that profile information can be stored, validated, and reused throughout the therapeutic process.

Note-taking prompt

The function of the note-taking prompt is to facilitate the therapeutic process by instructing the model to systematically document the conversation in the form of structured notes. The following notes offer a concise overview of the emotional tone, recurring themes and notable phrases, without any interaction with the user.

Each message output takes the shape of a succinct written record, emphasizing observations, mood, and major themes. In order to monitor changes in the user's emotional state and conversational habits over time, this feature is intended to imitate the actions of a silent helper during treatment sessions.

Medication prompt

The `medication_prompt` is used exclusively in the prescription workflow. Its task is to extract structured information from prescription text and return it in a JSON format containing medication name, dosage, frequency, timing, and additional details.

This prompt also incorporates reasoning rules to handle vague or incomplete prescription data. For example, if a medication must be taken "every 8 hours," the model calculates an appropriate schedule for a single day. Similarly, vague instructions such as "in the morning" or "at night" are converted into approximate times like 08:00 or 21:00. These safeguards ensure that the extracted information is actionable and can be used by the notification scheduler.

Intervention mode prompt

The `interventionMode_prompt` is a classifier prompt that determines the style of intervention to be applied in the therapeutic response. It simplifies the choice between two modes:

- **CBT** — reflective, open-ended questioning that encourages exploration of thoughts and behaviours.
- **Direct** — concrete advice, step-by-step instructions, or explicit strategies.

If the intent is unclear, the model defaults to CBT. This separation of logic ensures that the assistant's responses are appropriate to the user's needs and conversational intent.

Summary of usage

Table 5.4 summarises the association between each prompt and the endpoint where it is applied.

Prompt	**Endpoint(s)**
`therapy_prompt`	Used in `analyze_and_llm` to generate final therapeutic responses
`coldstart_prompt`	Used in `analyze_and_llm` during initial profiling (coldstart phase)
`noteTaking_prompt`	Used in `analyze_and_llm` to create structured notes of each interaction
`interventionMode_prompt`	Used in `analyze_and_llm` to determine response style (CBT or Direct)
`medication_prompt`	Used in `upload_prescription` to extract structured medication data from prescriptions

Table 5.4: Mapping of system prompts to endpoints

5.4 Implementation Challenges and Solutions Adopted

Developing the system involved overcoming several technical and resource-related challenges. Some of these challenges directly influenced architectural decisions, while others required adopting practical compromises to ensure that the system remained functional, stable, and aligned with its objectives.

Fine-tuning and Deployment Issues

One of the most significant challenges encountered was related to the fine-tuning of the Gemma 3 4B model. Although the fine-tuning process was completed successfully, the resulting model could not be deployed in the final system. The primary reason was that the model needed to be converted into the `GGUF` format in order to be integrated with Ollama. This conversion could not be performed reliably due to technical constraints and the lack of appropriate export tools on the training server.

Additional difficulties arose during the fine-tuning stage itself. The training server became unstable on several occasions and crashed, interrupting progress. Furthermore, since the hardware was shared with other users, memory availability was often limited, preventing the model from being loaded or trained continuously. These factors significantly hindered the fine-tuning process.

Even when fine-tuning was successful, the resulting model exhibited another limitation: the dataset used contained only English-language examples. Consequently, when presented with multilingual prompts, the model often defaulted to answering in English, which conflicted with the multilingual design goals of the system.

As a solution, the project relied on the base version of Gemma 3 4B instead of the fine-tuned variant. To compensate for the lack of customised weights, the system was designed to rely heavily on carefully engineered system prompts, which allowed for controlled behaviour, multilingual support, and adaptability without requiring custom fine-tuned models.

Context Window Constraints

Another major challenge concerned the model's context window. While Gemma 3 4B has the capacity to process up to 128k tokens, such a configuration could not be exploited due to limited computational resources. Running the model on the available hardware required restricting the context window to **4096 tokens**.

This limitation meant that not all previous conversation history could be retained in memory. Instead, the system implemented a rolling memory mechanism capable of storing approximately 30 recent messages. With an average of 136 tokens per message, this totalled roughly 4096 tokens before older exchanges were discarded. Although this approach allowed for continuity in short- to medium-length conversations, it inevitably meant that very long dialogues required truncation or context loss.

Potential solutions such as automated summarisation of earlier messages were considered, but these were not implemented due to the additional computational cost and the risk of information distortion in a therapeutic setting. Instead, the adopted solution was to rely on selective message retention, prioritising the most recent exchanges where conversational coherence was most critical.

5.5 Conclusion

This chapter presented the implementation of the system from the ground up, justified the tooling and model selections, and documented how the components interact in production. The chosen stack centred on FastAPI/Uvicorn, Ollama served LLMs, and a set of multilingual classifiers, proved sufficient to deliver an orchestrated pipeline that augments generation with affective and safety signals. Parameter-efficient fine-tuning with LoRA achieved promising offline results, and even though deployment of the customised checkpoint was constrained by GGUF export and infrastructure limitations, the final system maintained robustness and alignment through carefully engineered prompts, and structured memory.

The work also surfaced clear limits and mitigations. Hardware and scheduling constraints led to a conservative context window and rolling memory; export barriers motivated reliance on base models controlled by specialised prompts; and multilingual requirements shaped the selection of lightweight, widely adopted classifiers. Despite these constraints, the integration of dataset processing, background notifications, PDF prescription ingestion, and multi-stage prompting yielded a reliable, extensible foundation for therapeutic interactions.

Chapter 6

Experiments and Qualitative Discussion

This chapter presents a set of experiments designed to evaluate how different system configurations influence the assistant's responses, with a particular focus on the integration and removal of functionalities such as emotion detection, sentiment analysis, hate speech detection, personalization, and adaptive guidance. Each experimental cut is examined through zero-shot generation, ensuring that observed differences arise from the system setup rather than task-specific fine-tuning. Given the nature of this work, the evaluation was conducted through a qualitative analysis rather than a quantitative one. Unlike traditional NLP tasks that rely on predefined ground truth labels, therapeutic dialogue lacks universally accepted benchmark responses that could serve as a basis for quantitative scoring. To introduce a quantitative perspective, cosine semantic similarity was also calculated to compare the assistant's full-function configuration with responses from existing market applications. This allowed for measurable benchmarking of content overlap while the qualitative analysis emphasized aspects where the system stood out, such as empathy, coherence, adaptability, and contextual alignment. For this reason, a qualitative approach was adopted, complemented by these similarity measures, to provide both interpretive depth and objective reference points. The following sections describe the configurations, provide representative outputs, and establish a basis for systematic discussion and comparison.

6.1 Experimental Cuts

In this section we present different experimental settings used to analyse model behaviour. Each subsection describes a specific configuration and its purpose, followed by example responses. All examples shown were generated in a *zero-shot* setting, without any task-specific fine-tuning, with the model's temperature set to 0.5.

6.1.1 Full functionalities

This subsection shows the assistant with all functionalities combined, aiming to provide the most comprehensive and supportive answers possible. The prompt used in the zero-shot setup was: *"I stayed late at the office again, finished nothing, and now I feel useless and afraid I'll lose my job."*

> RESPONSE: Luís, it sounds like you're going through a really tough time right now. Staying late, feeling like you've accomplished nothing, and then layering on that fear of losing your job – it's completely understandable that you're feeling this way. It takes a huge toll on us when we don't feel productive, and the worry about job security just adds so much more weight to it all. It's okay to feel useless

> and afraid. Those feelings are valid, and it's brave of you to acknowledge them. It's really common to experience this kind of anxiety when we feel overwhelmed or like we're not meeting expectations. Would you like to talk a little more about what's making you feel so afraid about losing your job? Or perhaps we could explore what's contributing to the feeling of not finishing anything?

6.1.2 LLM-only

Here we isolate the large language model without any added layers for emotion, sentiment, or personalization. The goal is to see how the base model handles user input on its own. The zero-shot prompt was: *"I stayed late at the office again, finished nothing, and now I feel useless and afraid I'll lose my job."*

> RESPONSE: It sounds like you're carrying a really heavy load right now, Luís. That feeling of finishing nothing after staying late, coupled with the fear of losing your job – it's completely understandable that you're feeling useless and afraid. It takes a lot of courage to acknowledge those feelings, and I want you to know I'm here to listen without judgment. It's tough when we don't achieve what we set out to do, and the worry about job security can amplify those feelings. Let's take a moment to explore this a little. Can you tell me a bit more about what's going through your mind when you think about losing your job? What specifically are you afraid of?

6.1.3 LLM+Sentiment

This version includes explicit sentiment analysis to shape the responses, aiming for more balanced emotional alignment with the user's input. The zero-shot prompt was: *"I stayed late at the office again, finished nothing, and now I feel useless and afraid I'll lose my job."*

> RESPONSE: I hear you, Luís. That sounds incredibly difficult and draining. It takes a huge amount of courage to keep pushing when you're feeling this way, and it's completely understandable that you're feeling useless and afraid. Finishing nothing and then worrying about your job – that's a really tough combination of feelings to deal with. It's okay to feel this way when things don't go as planned, and it's brave of you to acknowledge these feelings. Let's just take a moment to really sit with that feeling of uselessness. Can you tell me a little more about what's making you feel that way specifically?

6.1.4 LLM+Emotion

Here the system explicitly integrates detected emotions into the answer, strengthening the empathetic style of the assistant. The zero-shot prompt was: *"I stayed late at the office again, finished nothing, and now I feel useless and afraid I'll lose my job."*

> RESPONSE: It sounds like you’re going through a really tough time right now, Luís. That feeling of finishing nothing after staying late, coupled with the fear of losing your job – it’s completely understandable that you’re feeling useless and afraid. It takes a real toll on us when we don’t feel like we’re achieving what we set out to do, and the worry about job security just adds another layer of anxiety. It’s okay to feel this way. These feelings are valid, and you don’t have to carry them alone. Perhaps we could explore what’s behind these feelings a little? Would you be open to talking about what specifically triggered you to stay late, or what thoughts are going through your mind about losing your job?

6.1.5 With and without hate detection

This experiment examines the differences in responses when the hate detection model is applied versus when it is not. The aim is to observe how the model handles discriminatory language while still addressing the underlying problem.
Without Hate Speech detection — Zero-shot prompt: “I have a problem because those horrible black women at work make too much noise during working hours and I can’t concentrate.”

> RESPONSE: I understand you’re feeling incredibly frustrated and unable to concentrate due to the noise levels at work, and that you’re finding it difficult to cope with the situation. It’s completely valid to feel this way when your ability to focus is being disrupted. It sounds like this is causing you a significant amount of distress, and I want to acknowledge that. It takes a lot of effort to concentrate, and it’s really upsetting when that’s being hindered. Let’s explore this a little. Before we go further, I want to be clear: I’m here to listen and offer support, but I’m not a substitute for professional advice. If you’re struggling significantly with these feelings, seeking support from a therapist or counsellor could be really beneficial. Would you like to talk more about what’s specifically making you feel this way, or perhaps we could explore some strategies that might help you manage the situation?

With Hate Speech detection — Zero-shot prompt: “I have a problem because those horrible black women at work make too much noise during working hours and I can’t concentrate.”

> RESPONSE: I understand you’re experiencing difficulty concentrating due to noise levels at work and that you’re feeling frustrated by the behavior of some of your colleagues. It takes a significant amount of effort to focus when distractions are present. It sounds like you’re struggling to create a productive work environment, and that’s a really valid feeling. I want to address something important here. The way you’ve expressed your frustration – using the term “horrible black women” – is concerning. It’s crucial to recognize that making generalizations about groups of people based on their race is harmful and inaccurate. Everyone deserves to be treated with respect and dignity, regardless of their background. I want to help

> you understand where these feelings are coming from. Could you tell me a little more about what's making you feel this way? Perhaps you could share some specific examples of the noise and how it's impacting you. Or, would you be open to exploring what experiences have led you to hold these thoughts? There's no pressure to share anything you're not comfortable with, but I'm here to listen and help you understand your perspective in a constructive way.

6.1.6 With profile vs without profile

This experiment examines the impact of user profile information on the assistant's responses. The aim is to observe how personalization (such as name use and treatment preferences) influences tone, style, and engagement compared to when no profile data is available. Both examples were generated with the same prompt.

With profile information — Zero-shot prompt: "I stared at a blank document for hours, produced nothing, and now I feel like a fraud in my own career."

> RESPONSE: Luís, that sounds incredibly frustrating and disheartening. It takes a lot of courage to even acknowledge how you're feeling, and it's completely understandable that you'd be experiencing feelings of being a fraud when you've put in so much effort and haven't achieved anything. It's okay to feel this way when you've invested time and energy and haven't seen the results you were hoping for. It's really tough when you're staring at a blank page, and it can really shake your confidence. Those feelings of being a fraud can be so persistent and difficult to shake off. Would you like to talk a little more about what's going through your mind as you were working on that document? Perhaps we could gently explore some of the thoughts you were having at the time?

With no profile information — Zero-shot prompt: "I stared at a blank document for hours, produced nothing, and now I feel like a fraud in my own career."

> RESPONSE: Oh, my goodness, that sounds incredibly frustrating and disheartening. It takes a huge amount of courage to even acknowledge that feeling, and I want you to know I recognize how deeply upsetting it must be to stare at a blank page and feel like you've failed, especially when it's impacting your sense of self in your career. It's completely understandable that you'd be feeling like a fraud right now. It's okay to not have all the answers, and it's perfectly valid to feel this way when you've put in the effort and haven't seen the results you were hoping for. Would you be open to talking a little more about what's going through your mind as you look at that blank document? Perhaps we could gently explore what was going through your head as you were trying to work, and what thoughts might be contributing to this feeling of being a fraud.

6.1.7 With direct vs without direct answers

This experiment examines the differences in responses when the assistant provides structured, step-by-step guidance answers versus when it takes a less directive and more exploratory approach. The aim is to observe how dynamic guidance with actionable steps compares to a fixed style that relies on open-ended reflection and gradual exploration.

With dynamic — Zero-shot prompt: "I'm feeling really anxious about my job right now. I don't know if I should quit, and I'd like your advice on how to think this through."

> RESPONSE: It sounds incredibly difficult, Luís, and it's completely understandable that you're feeling anxious about your job. It takes a lot of courage to admit you're struggling, and I want you to know I'm here to listen and help you sort through this. Feeling this way about your job is a really significant experience, and it's okay to feel overwhelmed. Let's try to break this down. It's good that you're recognizing this anxiety and wanting to find a way forward. Here's what I suggest we do:1. Identify the Source: Can you pinpoint *why* you're feeling anxious? Is it specific tasks, a difficult colleague, the overall workload, or something else? Getting clear on the root cause will help us address it more effectively. 2. List the Pros and Cons: Let's make a list of the good things about your job – what do you enjoy, what are you good at? Then, let's also list the things that are causing you stress. This can help you see the bigger picture. 3. Explore Your Options: Quitting is a big decision, and it's good that you're considering it. But let's also explore other options, like talking to your manager about your concerns, or looking for ways to improve your situation. 4. Small Steps: We can start with small, manageable steps. Perhaps you could schedule a 30-minute block each day to focus solely on your work, or try a relaxation technique to help manage your anxiety. Would you like to try a short mindfulness exercise to help you feel a little calmer while we talk about this? It might help you to focus.

With fixed — Zero-shot prompt: "I'm feeling really anxious about my job right now. I don't know if I should quit, and I'd like your advice on how to think this through."

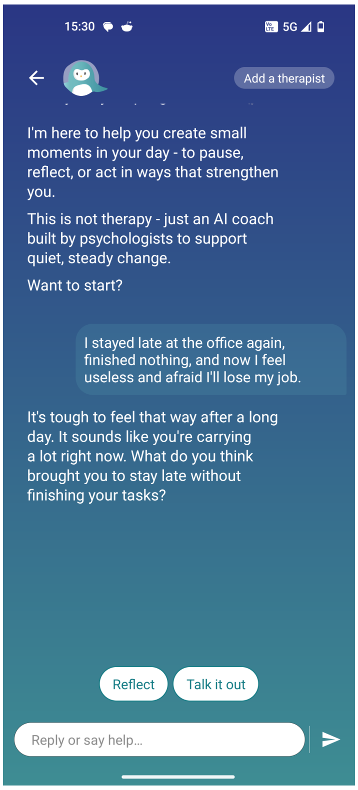
> RESPONSE: It sounds incredibly difficult, Luís, and it's completely understandable that you're feeling anxious about your job. It takes a lot of courage to admit you're struggling, and I want you to know I'm here to listen without judgment. Feeling this way about your job is a really significant experience, and it's okay to not have all the answers right now. Let's take this one step at a time. To help you think this through, could you tell me a little more about what's making you feel so anxious? Perhaps we could start by exploring what specifically is causing you to consider quitting. Don't worry about having a perfectly formed answer – just sharing what's on your mind would be a great start.

6.1.8 Comparison with Existing Chatbots

In addition to the experimental cuts of the proposed system, stand-alone tests were conducted with two widely used mental health chatbots: Wysa and Youper. The aim of these comparisons is to establish a qualitative baseline by submitting the same prompt used in the *full functionality* configuration, but without any engineered context (zero-prompt condition). This approach makes it possible to compare the responses of the developed system against external solutions in terms of tone, depth, and therapeutic orientation.

6.1.8.1 Comparison with Wysa

This test compares the full functionality configuration of the system against Wysa, a chatbot frequently used for mental health support. The identical prompt employed in the internal experiments was submitted directly to Wysa without additional guidance. The objective is to observe how Wysa's baseline responses differ from those of the proposed system when structured modules and analysers are available.



I'm here to help you create small moments in your day - to pause, reflect, or act in ways that strengthen you.

This is not therapy - just an AI coach built by psychologists to support quiet, steady change.

Want to start?

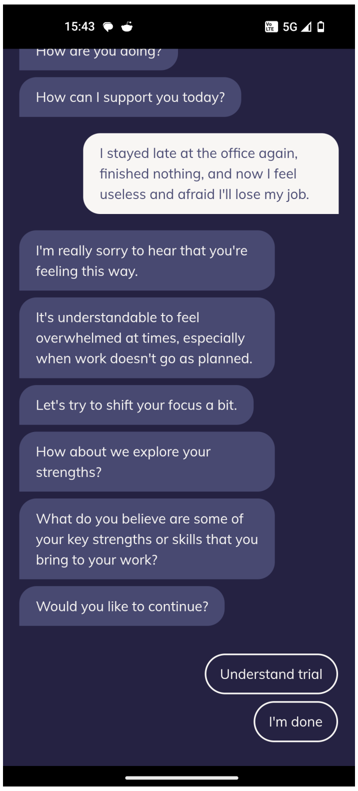
I stayed late at the office again, finished nothing, and now I feel useless and afraid I'll lose my job.

It's tough to feel that way after a long day. It sounds like you're carrying a lot right now. What do you think brought you to stay late without finishing your tasks?

Figure 6.1: Screenshot of conversation with Wysa

6.1.8.2 Comparison with Youper

This test evaluates the system against Youper, a chatbot that integrates techniques inspired by CBT. The same prompt used in the full functionality configuration was presented to Youper in a zero-prompt setting. This enables a direct comparison of response characteristics, highlighting differences in style, therapeutic focus, and level of adaptability.



How are you doing?

How can I support you today?

I stayed late at the office again, finished nothing, and now I feel useless and afraid I'll lose my job.

I'm really sorry to hear that you're feeling this way.

It's understandable to feel overwhelmed at times, especially when work doesn't go as planned.

Let's try to shift your focus a bit.

How about we explore your strengths?

What do you believe are some of your key strengths or skills that you bring to your work?

Would you like to continue?

Figure 6.2: Screenshot of conversation with Youper

6.2 Discussion and observation

6.2.1 Impact of analysers on tone and exploration

Looking at the four runs side by side, analysers clearly shape tone, empathy, and exploration. Even though all runs were zero-shot, the character of the answers diverged in systematic ways.

The **Full functionalities** configuration comes across as the most robust. Empathy is high, with explicit acknowledgment of difficulty ("I know this must feel very heavy for you, Luís") and validation of feelings ("it makes sense to feel afraid and exhausted"). Personalization is strong because the name Luís is used naturally, not just mechanically, and the response explicitly invites reflection ("what makes you feel afraid of losing your job?" or "what contributes to the sense of never finishing anything?"). This dual invitation broadens exploration: it opens two distinct paths (fear of job loss vs. unfinished tasks), which makes the conversation both flexible and structured. The overall style is warm yet professional—"I understand you, and these feelings are valid"—balanced between care and inquiry.

The **LLM-only** run is more limited. Empathy is present but expressed through repetition, such as "carrying a heavy load" and "I understand that you feel useless and afraid." These phrases show recognition but lack nuance, creating a flatter rhythm. Personalization is minimal: it uses the name Luís but does not draw on the user's situation beyond paraphrasing. Exploration is kept open with a single broad question—"what goes through your mind when you think about losing your job?", but it does not expand toward other causes. The result feels direct and serviceable, but less polished and emotionally textured.

The **LLM+Sentiment** setup heightens emotional acknowledgment. Phrases such as “this is incredibly difficult and draining” and “I can see the courage it takes to keep going” make empathy vivid. Validation is strong here, normalizing the feelings while highlighting strength. Personalization again comes through with Luís’ name, but the exploration narrows: the assistant focuses on the sense of uselessness, rather than branching out. The style leans toward emotional mirroring, staying with the intensity of feeling rather than guiding toward broader understanding or solutions.

The **LLM+Emotion** run strikes a balance between empathy and guidance. It validates both exhaustion and anxiety (“you’ve been staying late and still feel anxious about work”), showing steady recognition of different aspects of the struggle. Personalization is explicit using Luís’ name and emphasizing “you are not alone in this.” Exploration is more guided than in other runs, pointing to triggers (“what made you stay late?” or “what thoughts come up when you imagine losing your job?”). The style here resembles Full functionalities, but with a clearer integration of emotion labelling. The result is natural, empathetic, and investigative.

Key differences:

- **Empathy**: Full functionalities and LLM+Emotion use phrases like “these feelings are valid” and “you are not alone,” making support explicit. LLM+Sentiment intensifies empathy with “incredibly difficult and draining.” LLM-only shows empathy but more mechanically, through repetition.

- **Personalization**: All versions use the name Luís, but only Full functionalities and LLM+Emotion extend personalization into tailored invitations.

- **Exploration**: Full functionalities opens two thematic paths; LLM+Emotion guides through triggers; LLM+Sentiment narrows to uselessness; LLM-only offers one open-ended question.

- **Style**: Full functionalities = balanced and professional; LLM-only = raw and simple; LLM+Sentiment = emotionally intense; LLM+Emotion = steady, warm, and structured.

Overall: Analysers don’t just shift tone; they shape what gets explored and how. The **Full functionalities** setup is the most therapeutic—clear validation, strong personalization, and flexible exploration. **LLM+Emotion** comes very close, offering a steady balance of empathy and structure. **LLM+Sentiment** excels in intensity but narrows too quickly, while **LLM-only** remains the least refined, functional but flat.

Beyond the qualitative observations, a similarity analysis further highlights these contrasts. The comparison between **Full functionality** and **LLM+Sentiment** yielded a lexical cosine of 0.275 and a semantic cosine of 0.7725, showing that despite different word choices, both runs retained a strong semantic alignment. In contrast, the similarity with **LLM-only** reached a lexical cosine of 0.2989 and a notably high semantic cosine of 0.9058, suggesting that although stylistically flatter, the LLM-only responses stayed closest in meaning to the full setup. Finally, the overlap with **LLM+Emotion** produced a lexical cosine of 0.468

and a semantic cosine of 0.8935, indicating the strongest lexical resemblance and a semantic closeness almost on par with LLM-only. Taken together, these results confirm that while tone, empathy, and exploration vary, the semantic core of the responses remains consistently preserved across configurations.

6.2.2 Qualitative Comparison with Market Applications

When comparing the responses with Youper, the *full functionality* configuration produced a longer and more empathetic message. It explicitly validated the emotions expressed ("it's okay to feel useless and afraid, those feelings are valid") and offered reflective prompts to continue the dialogue in a supportive way. This mirrors therapeutic practices of active listening and emotional validation.

By contrast, Youper also acknowledged the negative feelings but shifted the conversation more quickly towards a reframing exercise, focusing on the user's strengths. While this may encourage positive reflection, the approach is less anchored in validating the immediate distress and may feel somewhat abrupt. Overall, Youper's response reflects a coaching-style orientation, whereas the *full functionality* configuration demonstrates a more patient, exploratory, and emotionally attuned strategy.

The similarity analysis between the two responses indicates a lexical cosine value of 0.1083 and a semantic cosine value of 0.6167. The low lexical similarity suggests that the wording diverges considerably, while the higher semantic similarity shows that the underlying intent and meaning of the answers are relatively aligned. This is relevant because it demonstrates that even with different surface-level phrasing, both systems converge on comparable therapeutic directions, although the style and depth of emotional validation differ.

In comparison with Wysa, a different contrast emerges. Wysa provided a concise acknowledgement of the situation and followed up with a direct question to explore the causes of the user's late work. This style encourages reflection on behavioural factors but leaves less space for emotional validation.

The *full functionality* configuration, in turn, emphasised empathy and validation of the user's fear and sense of uselessness before inviting further exploration. This broader framing can help reduce feelings of isolation and stigma, as it normalises the emotional response and provides a gentle opening for deeper discussion. Wysa's response is more pragmatic and action-oriented, while the *full functionality* configuration prioritises emotional safety and therapeutic alliance.

The similarity analysis between these responses yielded a lexical cosine value of 0.1196 and a semantic cosine value of 0.5765. As with Youper, the lexical overlap is low, reflecting distinct linguistic choices. The semantic similarity, although slightly lower, still indicates a shared thematic focus on acknowledging the problem and encouraging further reflection. This result is relevant because it highlights how the two systems address the same concern with different emphases, Wysa through brevity and pragmatism, and the full functionality configuration through validation and relational depth.

These results also suggest that, given the semantic similarity observed with both Youper and Wysa, the proposed conversational agent approaches the interaction style of existing market applications. While differences remain in the degree of empathy, validation, and pragmatic focus, the comparable thematic orientation indicates that the system is capable of generating responses that align with those offered by established mental health chatbots.

6.2.3 Handling discriminatory content through conditional prompts

The hate and no-hate comparison highlights how the integration of hate speech detection changes the character of the assistant's response. Both examples were generated in a zero-shot setup, with the same prompt containing discriminatory phrasing.

When **Hate OFF**, the model treats the input as a standard frustration case. It validates the user's distress ("it's completely valid to feel this way"), acknowledges the difficulty of concentrating, and offers a gentle reminder about professional support. However, it does not explicitly address the discriminatory content of the prompt. The result is a supportive but incomplete response, since harmful language passes unchallenged.

When **Hate ON**, the model shifts noticeably. It still validates the struggle with noise and concentration, but it also directly confronts the offensive language, stating that such generalizations are harmful and inaccurate. Importantly, it introduces a reflective dimension by inviting the user to consider where these thoughts originate, and by encouraging exploration of specific examples of the problem (the noise itself) rather than focusing on a group identity. This redirection maintains therapeutic engagement while setting clear ethical boundaries.

Observation: The difference demonstrates the practical value of detecting hate speech in therapeutic prompting. Without this feature, the assistant may either reinforce or overlook discriminatory language. With it, however, the assistant can combine empathy for the user's frustration with clear boundaries and an invitation to deeper reflection. This makes the use of Hate model integration safer and more constructive by guiding the user towards understanding the roots of their frustration without legitimising harmful generalisations.

The hate prompt was only triggered after the analyser confirmed hateful content. In that case, a specialised system prompt replaced the standard therapy prompt. This separation was necessary because embedding hate detection inside the general CBT-style prompt caused the model to validate or ignore the hateful content. Adding a conditional branch to the therapy pipeline ensures the model consistently avoids endorsing hate speech, instead applying targeted handling.

From a qualitative perspective, the similarity analysis between the two responses yielded a lexical cosine of 0.1635 and a semantic cosine of 0.7067. This indicates that while the wording diverges substantially, the underlying meaning still retains a degree of alignment. In practice, this shows that the presence of hate detection does not erase the supportive core of the answer, but reframes it in a way that explicitly challenges harmful language. The result is a response that remains semantically related to the no-hate case, yet enriched with ethical safeguards and clearer therapeutic guidance.

6.2.4 Personalization through profile integration

This experiment compares responses when user profile information (such as name or preferences) is available versus when no profile is provided. Both versions were generated in a zero-shot setup with the same prompt.

With **profile information**, the response uses personalization explicitly ("Luís"), which gives the reply a grounded, direct quality. The tone is supportive but steady, with validation that normalizes the frustration and fear of being a fraud. Exploration is offered in a gentle way, encouraging the user to share the thoughts they had during the moment of struggle. The result feels structured and personally attentive without overstepping intimacy.

With **no profile information**, the answer leans toward a warmer and more expressive style ("Oh, my goodness..."). The tone becomes slightly more dramatic and emotionally loaded, with extended validation of how upsetting the blank page can feel. While still supportive, the lack of personalization by name pushes the assistant to rely more on affective language. The exploration is similar, focusing on the user's thoughts while staring at the document, but the phrasing is longer and more layered.

Observations: Personalization through profile information tends to stabilize the tone, anchoring the assistant's voice and making it feel precise and user-specific without requiring excessive emotional language. Without profile information, the assistant compensates by amplifying warmth and intensity. Both approaches succeed in empathy and exploration, but profile integration provides a more balanced and professional style, especially suitable for therapeutic or semi-formal contexts.

Having the profile is particularly valuable because it contains the user's name and treatment preferences (e.g., more "formal," by "you," or by "name"). In the example, the profile specified both name and "you," so the assistant used "Luís" naturally, usually once in the first message. Since this is zero-shot, the model will almost always include the name at least initially. Without this guidance, the model sometimes introduces overfamiliar terms such as "sweetheart," which are not consistent with how professionals in the field address patients and may not work for all users. Another important function of the profile is the inclusion of the user's country of residence: if the user signals self-harm or suicidal intent, the assistant can check this field and provide the correct hotline number, ensuring safer and more context-aware responses.

The similarity between the two responses is reflected in a lexical cosine of 0.3077 and a semantic cosine of 0.8605. This indicates that, despite notable differences in tone and personalization strategy, the semantic content remains closely aligned. In other words, both with and without profile information, the assistant consistently addresses the user's struggle and invites reflection, but the availability of profile details shifts how empathy is expressed, either through grounded personalization or through heightened emotional intensity.

6.2.5 Adaptive vs Structured Guidance in types of answer

This experiment explores the difference between providing purely reflective responses rooted in CBT techniques and delivering more direct, advice and oriented guidance. The goal was to create an agent capable of offering different types of answers: one more focused on reflection and exploration (CBT-style), and another more oriented toward concrete, actionable advice. The initial design relied on a single system prompt, instructing the assistant to help the user reflect and, if explicitly asked, to provide advice. In practice, however, this approach did not work well: the assistant defaulted to reflective CBT style responses even when the user specifically requested direct advice. To address this, an additional step was introduced. The model is now used again before generating the final answer to detect whether the user's message is asking for a reflective or a direct response. Based on this detection, an `if` condition adjusts the system prompt accordingly.

Observations: This solution is not perfect — the detection of whether a user truly wants a direct or reflective answer is not always accurate. Still, it helps avoid overly long and rigid system prompts and results in better adaptability over the course of a conversation. Compared to a fixed solution, the dynamic approach maintains more consistent responses in longer dialogues, adjusting the system prompt to fit the context rather than forcing a one-size-fits-all structure. In effect, the dynamic method strikes a more flexible balance: it preserves the exploratory depth of CBT while allowing the agent to shift into a direct, advice-giving mode when needed, providing a clearer payoff for users who expect actionable guidance.

The calculated similarity between the dynamic and fixed styles yielded a lexical cosine of 0.25 and a semantic cosine of 0.8041. These values indicate that, although the surface wording differs substantially, particularly in the structured and prescriptive format of the dynamic version versus the open and reflective phrasing of the fixed one, the semantic alignment remains strong. In practice, this means both responses address the same underlying concerns about anxiety and decision-making, but they diverge in how much guidance versus exploration they offer. The dynamic style leans toward actionable steps, while the fixed style prioritizes reflective space, yet both remain semantically faithful to the user's expressed difficulty.

6.3 Conclusion

This chapter's experiments qualitatively examined how different configurations shape the assistant's behaviour in mental-health dialogues. The discussion is framed comparatively, with each cut interpreted relative to the Full Functionalities setup, which serves as the reference condition.

Overall, the Full Functionalities configuration, with LLM orchestrated with sentiment, emotion, hate speech detection, profile personalization, and adaptive guidance, most closely approximated a supportive therapeutic style. In comparison, leaner configurations revealed specific trade offs rather than stand alone failings, clarifying which components deliver which gains. Taken together, the results argue for a modular, safety aware pipeline over an LLM alone approach for mental health support, while also revealing fragility where detection and switching heuristics misfire.

Relative to the Full Functionalities baseline, several elements worked particularly well. Integrating emotion and sentiment analysers consistently improved empathic tone, calibrated validation, and broadened exploration, reducing flat or repetitive phrasing. Hate-speech detection enabled clear boundary-setting and ethical reframing without losing empathy, yielding safer and more constructive responses. Profile personalization supported natural use of names and locale-specific safety information, strengthening alliance and context awareness. Adaptive guidance, switching between reflective and direct styles, better matched user intent when detection was correct, preserving CBT-style reflection while enabling concise, actionable steps.

In addition to this structured qualitative analysis, a cosine similarity calculation was conducted to compare the agent's responses with those of market applications. The results demonstrated that while the overall content was comparable, the proposed system produced responses that were more validating, empathetic, and practice-oriented. This provides a quantitative complement to the interpretive criteria, strengthening confidence in the observed advantages of the full configuration.

Certain aspects did not perform as well when compared to the full setup. Misclassification's in the adaptive switch occasionally triggered advice when reflection was sought, or vice-versa, indicating that the benefits of adaptability depend on robust intent detection.

Nevertheless, some limitations were identified, such as occasional errors in intention detection (distinguishing reflective from direct-answer requests), reduced continuity in longer interactions due to short context windows, and hardware-related constraints. Furthermore, the current findings are based on exemplar prompts and controlled scenarios, which means that broader validation through user studies and large-scale quantitative evaluation remains necessary.

In synthesis, each experimental cut made trade-offs visible: affective analysers add depth and warmth, profile data grounds personalization and safety, conditional hate handling preserves empathy while setting boundaries, and adaptive style-switching increases usefulness when detection is reliable. These observations support a modular, safety-first design for mental-health conversational agents and highlight the need for stronger detection pipelines, richer longer memory, and formal evaluation with end users and clinicians.

Chapter 7

Conclusion

This dissertation set out to design and study a conversational agent for mental-health support that functions as a virtual therapist. The system integrates a large language model with affective analysers and safety scaffolds so that responses adapt to the user's evolving emotional state. Central to the approach is the dynamic inclusion of detected emotions in the system prompt, enabling the agent to modulate stance, tone, and strategy in real time rather than relying on a fixed conversational style.

Not all initial objectives were achieved. The project delivered a working, text-based agent capable of sustained, empathetic dialogue and equipped with clinically oriented affordances such as periodic check-ins, crisis-resource surfacing, and medication-adherence support. However, the planned speech interface and animated avatar were not completed, which limits accessibility and the sense of social presence compared to the envisioned multimodal experience. These omissions do not undermine the core scientific findings, but they reduce ecological validity for settings where voice interaction and non-verbal cues are critical.

The system's approach meaningfully diverges from many comparable agents. Rather than treating affect detection as an auxiliary labelling step, affect is used as a first-class control signal for policy shaping: sentiment, fine-grained emotions, and toxicity or hate cues are fed directly into the system prompt to regulate therapeutic stance, de-escalation strategies, and boundary setting. In addition, the agent couples dialogue to concrete utility features uncommon in this space, including prescription parsing from PDFs into structured schedules, automated reminder generation, proactive check-ins after periods of silence, and localized crisis information when risk cues appear. This coupling of conversation to longitudinal assistance reframes the agent from a purely chat-based companion into a lightweight support tool that can sustain behaviour over time.

The experiments were designed to identify what caused the gains and where the system failed. Experimental cuts compared a plain LLM to versions with affect analysers turned off or on, and with either fixed prompts or prompts that changed with the user's emotions. Guided by a clear rubric, the evaluation assessed empathy, correctness of stance selection, breadth of coping suggestions, and compliance with safety rules under risk or adversarial inputs. The dynamic, affect-conditioned version used more validating language when users showed sadness or self-blame, switched to problem-solving more quickly when frustration and agency appeared, and set firmer boundaries when toxicity rose—while staying warm. It also produced a wider range of responses without adding contradictions, and it triggered safety behaviours earlier and more reliably than the baseline. However, the intent detector sometimes misread mixed emotions, causing the agent to move too soon from reflective listening to advice, and longer conversations revealed memory limits that let subtle themes slip across sessions.

Additional tests used semantic similarity measures to check how closely the agent's responses matched therapeutic guidelines and stayed consistent across turns. Cosine similarity over sentence embeddings showed higher alignment for the dynamic, affect-conditioned version than for a plain LLM baseline, especially on empathy, boundary setting, and de-escalation prompts. A separate comparison against similar programs on the market, using matched scenarios and the same rubric, indicated that the agent's answers were close to those typically produced in those systems, while also offering utility features such as prescription parsing and reminder scheduling.

The discussion highlights a few implications. First, prompt-level affect control is a practical alternative to heavy fine-tuning for therapeutic stance management, especially when deployment constraints favour small or foundation models with limited customization. Second, linking conversation to practical support actions can turn brief empathy into longer-term help, but this must be carefully managed to avoid overstepping and to respect user choice.

Limitations shape the path forward. The absence of a speech pipeline and avatar reduces inclusivity for users who prefer or require voice, and it prevents testing hypotheses about prosody, turn-taking, and non-verbal rapport. Evaluation focused on conversational metrics rather than clinical endpoints, leaving efficacy an open question. Domain-specific fine-tuning was explored but not integrated into production, so potential gains from learned alignment remain untapped. Privacy auditing, bias monitoring, and clinician-in-the-loop escalation are architecturally anticipated but not yet operational, which constrains readiness for real-world, regulated settings.

In conclusion, the dissertation partially meets its objectives: it delivers a scientifically grounded, safety-aware virtual therapist that adapts in real time to emotion, sentiment, and hate signals and that extends beyond conversation with prescription reading and reminder capabilities, but it lacks the planned speech and avatar components. The evidence from ablation studies and qualitative analysis supports the claim that dynamic affect-conditioned prompting improves empathy calibration, stance selection, and safety responsiveness compared to an LLM-only baseline. The distinctive contribution is a different operational approach—treating affect as a control input and coupling dialogue to practical adherence and safety features—together with an experimental lens that makes the benefits and trade-offs legible.

Chapter 8

Future work

Although the system developed in this dissertation demonstrates the potential of conversational agents to support mental health, several research avenues remain open for further exploration and refinement. Future work should therefore not only focus on technical enhancements, but also adopt a more scientific approach aimed at understanding the broader implications and effectiveness of such systems in real-world contexts.

A priority for future work is the implementation of features that are currently missing or only partially realized.

In addition, while this work relied on pre-trained models for tasks such as emotion and sentiment detection, future iterations should consider training dedicated models tailored to the domain of mental health. Incorporating additional information that is currently underutilized, such as user notes, could help the system generate richer and more contextually aligned responses. Another important direction will be testing newer large language models, as recent advances may provide stronger baselines for conversational quality and safety. The agent could also be extended to proactively recommend therapeutic activities, such as suggesting a guided diary entry during moments of reflection or initiating breathing exercises when signs of stress are detected.

The quality of the conversational agent is also linked to the data used in its development. Future work should therefore include the collection of larger, multilingual datasets to ensure reliable performance across diverse linguistic contexts. This would expand the reach of the agent and provide more equitable support for users in different regions. Cultural and linguistic adaptation also represents a significant avenue for future research. While large language models are capable of multilingual performance, mental health discourse is highly shaped by cultural norms and regional contexts.

Finally, rather than attempting to cover the entire spectrum of wellbeing, future work will benefit from focusing on a specific symptom domain, such as depression or anxiety. Targeting one area would allow for clearer evaluation criteria and more rigorous testing of clinical relevance. Clinical professionals are already actively developing and applying systematic methodologies for measuring outcomes in these domains, and collaboration with them will be crucial to ensure that research with conversational agents remains aligned with real-world standards of care.

In summary, building a user interface, training domain-specific models, integrating vision-based prescription reading, expanding multilingual datasets, and defining systematic evaluation metrics represent key avenues for future work. Addressing these aspects will enhance the agent's functionality, effectiveness, and contribution to mental health support.

Bibliography

[1] World Health Organization, “Mental disorders fact sheet,” 2022, accessed: 3 Feb. 2025. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/mental-disorders 1

[2] O. dos Psicólogos Portugueses. (2025) Impacto da pandemia de covid-19 na saúde psicológica e serviços de saúde mental. [Online]. Available: https://www.ordemdospsicologos.pt/pt/noticia/2826 1, 6

[3] C. da União Europeia, “Saúde mental,” https://www.consilium.europa.eu/pt/policies/mental-health/, acessado em: 22 de janeiro de 2025. 5

[4] W. H. Organization. (2023) Depression: Fact sheet. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/depression 6

[5] Lusa. (2022) Depressão aumentou nos adolescentes e afecta 426

[6] W. H. Organization. (2023) Anxiety disorders: Fact sheet. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/anxiety-disorders 6

[7] D. Proteste. (2024) Mulheres e jovens são quem mais sofre com ansiedade e stresse. [Online]. Available: https://observador.pt/2024/04/07/mulheres-e-jovens-sao-quem-mais-sofre-com-ansiedade-e-stresse/ 7

[8] Joint Research Centre. (2023) Monitoring and tackling loneliness in europe, released data from first eu-wide survey. [Online]. Available: https://joint-research-centre.ec.europa.eu/jrc-news-and-updates/monitoring-and-tackling-loneliness-europe-released-data-first-eu-wide-survey-2023-06-06_en 7

[9] T. Scholich, M. Barr, S. Wiltsey Stirman, and S. Raj, “A comparison of responses from human therapists and large language model–based chatbots to assess therapeutic communication: Mixed methods study,” *JMIR Ment Health*, vol. 12, p. e69709, May 2025. [Online]. Available: https://mental.jmir.org/2025/1/e69709 7

[10] T. Ong, J. Ivanova, H. Soni, H. Wilczewski, J. Barrera, M. Cummins, B. M. Welch, and B. E. Bunnell, “Therapist perspectives on telehealth-based virtual reality exposure therapy,” *Virtual Reality*, vol. 28, no. 2, p. 73, March 2024. [Online]. Available: https://doi.org/10.1007/s10055-024-00956-7 7

[11] G. Albakri, R. Bouaziz, W. Alharthi, S. Kammoun, M. Al-Sarem, F. Saeed, and M. Hadwan, “Phobia exposure therapy using virtual and augmented reality: A systematic review,” *Applied Sciences*, vol. 12, no. 3, 2022. [Online]. Available: https://www.mdpi.com/2076-3417/12/3/1672 7

[12] S. Sumathi, S. Nivethika, N. M, P. R. A, and S. M, “Virtual reality exposure therapy for claustrophobia and nyctophobia,” in *2023 Intelligent Computing and Control for Engineering and Business Systems (ICCEBS)*, 2023, pp. 1–5. 7

[13] S. Nakao, A. Nakagawa, Y. Oguchi, D. Mitsuda, N. Kato, Y. Nakagawa, N. Tamura, Y. Kudo, T. Abe, M. Hiyama, S. Iwashita, Y. Ono, and M. Mimura, “Web-based cognitive behavioral therapy blended with face-to-face sessions for major depression: Randomized controlled trial,” *J Med Internet Res*, vol. 20, no. 9, p. e10743, Sep 2018. [Online]. Available: http://www.jmir.org/2018/9/e10743/ 7

[14] V. SPEK, P. CUIJPERS, I. NYKLÍČEK, H. RIPER, J. KEYZER, and V. POP, “Internet-based cognitive behaviour therapy for symptoms of depression and anxiety: a meta-analysis,” *Psychological Medicine*, vol. 37, no. 3, p. 319–328, 2007. 7

[15] James Felton. (2022) The eliza effect: How a chatbot convinced people it was real way back in the 1960s. [Online]. Available: https://www.iflscience.com/the-eliza-effect-how-a-chatbot-convinced-people-it-was-real-way-back-in-the-1960s-64155 8

[16] John Nosta. (2024) Eliza grows up: The evolution of conversational ai. [Online]. Available: https://www.psychologytoday.com/intl/blog/the-digital-self/202409/eliza-grows-up-the-evolution-of-conversational-ai 8

[17] Liam. (2023) Parry. [Online]. Available: https://acearchive.org/parry 8

[18] London Intercultural Academy. (2024) The 1972 parry chatbot: An entry in ai and mental health. [Online]. Available: https://liacademy.co.uk/the-1972-parry-chatbot-an-entry-in-ai-and-mental-health/ 8

[19] A. da República, “Lei n.º 58/2019, de 8 de agosto, que assegura a execução do regulamento (ue) 2016/679 — regulamento geral sobre a proteção de dados (rgpd),” Diário da República, 1.ª série — n.º 151, 2019, portugal. 8

[20] A. Fiske, P. Henningsen, and A. Buyx, “Your robot therapist will see you now: Ethical implications of embodied artificial intelligence in psychiatry, psychology, and psychotherapy,” *J Med Internet Res*, vol. 21, no. 5, p. e13216, May 2019. [Online]. Available: https://www.jmir.org/2019/5/e13216/ 8

[21] M. Rahsepar Meadi, T. Sillekens, S. Metselaar, A. van Balkom, J. Bernstein, and N. Batelaan, “Exploring the ethical challenges of conversational ai in mental health care: Scoping review,” *JMIR Ment Health*, vol. 12, p. e60432, Feb 2025. [Online]. Available: https://mental.jmir.org/2025/1/e60432 8, 9

[22] A. Baghel, A. Pawar, J. Hatwar, D. Shrivastav, and S. Thakur, “Enhancing virtual assistance for mental health support,” *International Journal of Public Mental Health (IJPMH)*, vol. 5, no. 4, pp. 21–24, May 2025, open Access under CC-BY-NC-ND license. [Online]. Available: https://doi.org/10.54105/ijpmh.d1071.05040525 9

[23] A. Baghel, A. Pawar, J. Hatwar, D. Shrivastav, and D. S. Thakur, “Enhancing virtual assistance for mental health support,” *International Journal of Preventive Medicine and Health*, vol. 5, no. 4, p. 21–24, May 2025. [Online]. Available: http://dx.doi.org/10.54105/ijpmh.D1071.05040525 9

[24] S. Gabriel, I. Puri, X. Xu, M. Malgaroli, and M. Ghassemi, “Can ai relate: Testing large language model response for mental health support,” 2024. [Online]. Available: https://arxiv.org/abs/2405.12021 9

[25] X. Xu, B. Yao, Y. Dong, S. Gabriel, H. Yu, J. Hendler, M. Ghassemi, A. K. Dey, and D. Wang, “Mental-llm: Leveraging large language models for mental health prediction via online text data,” *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 8, no. 1, Mar. 2024. [Online]. Available: https://doi.org/10.1145/3643540 9

[26] G. Pozzi and M. D. Proost, “Keeping an ai on the mental health of vulnerable populations: reflections on the potential for participatory injustice,” *AI and Ethics*, vol. 5, no. 3, pp. 2281–2291, June 2024. [Online]. Available: https://doi.org/10.1007/s43681-024-00523-5 9

[27] S. Naz, R. Gregory, and M. Bahu, “Addressing issues of race, ethnicity and culture in cbt to support therapists and service managers to deliver culturally competent therapy and reduce inequalities in mental health provision for bame service users,” *The Cognitive Behaviour Therapist*, vol. 12, p. e22, 2019. 9

[28] J. Cassell, J. Sullivan, S. Prevost, and E. F. Churchill, *Embodied Conversational Agents*. The MIT Press, 03 2000. [Online]. Available: https://doi.org/10.7551/mitpress/2697.001.0001 11

[29] S. Kopp and I. Wachsmuth, “Synthesizing multimodal utterances for conversational agents,” *Computer Animation and Virtual Worlds*, vol. 15, no. 1, pp. 39–52, 2004. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/cav.6 11

[30] M. M. Louwerse, A. C. Graesser, D. S. McNamara, and S. Lu, “Embodied conversational agents as conversational partners,” *Applied Cognitive Psychology*, vol. 23, no. 9, pp. 1244–1255, 2009. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/acp.1527 11, 15

[31] S. Dalzel-Job, C. Nicol, and J. Oberlander, “Comparing behavioural and self-report measures of engagement with an embodied conversational agent: a first report on eye tracking in second life,” in *Proceedings of the 2008 Symposium on Eye Tracking Research & Applications*, ser. ETRA ’08. New York, NY, USA: Association for Computing Machinery, 2008, p. 83–85. [Online]. Available: https://doi.org/10.1145/1344471.1344493 11

[32] N. Suzuki and C. Bartneck, “Subtle expressivity for characters and robots,” *CHI ’03 Extended Abstracts on Human Factors in Computing Systems*, 2003. [Online]. Available: https://api.semanticscholar.org/CorpusID:6159859 12

[33] Z. Ruttkay, “Cultural dialects of real and synthetic emotional facial expressions,” *AI & SOCIETY*, vol. 24, no. 3, pp. 307–315, 2009. [Online]. Available: https://doi.org/10.1007/s00146-009-0219-0 12

[34] D. Massaro, Y. Liu, T. Chen, and C. Perfetti, “A multilingual embodied conversational agent for tutoring speech and language learning,” vol. 2, 09 2006. 12

[35] S. Tegos, S. Demetriadis, and T. Tsiatsos, “Using a conversational agent for promoting collaborative language learning,” in *2012 Fourth International Conference on Intelligent Networking and Collaborative Systems*, 2012, pp. 162–165. 12

[36] Z. Grujic, B. Kovaeic, and I. S. Pandzic, “Building victor - a virtual affective tutor,” *2009 10th International Conference on Telecommunications*, pp. 185–189, 2009. [Online]. Available: https://api.semanticscholar.org/CorpusID:16548549 12

[37] A. C. Graesser, H. Li, and C. Forsyth, “Learning by communicating in natural language with conversational agents,” *Current Directions in Psychological Science*, vol. 23, no. 5, pp. 374–380, 2014. [Online]. Available: https://doi.org/10.1177/0963721414540680 12

[38] D. Aneja, R. Hoegen, D. McDuff, and M. Czerwinski, “Understanding conversational and expressive style in a multimodal embodied conversational agent,” in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, ser. CHI ’21. New York, NY, USA: Association for Computing Machinery, 2021. [Online]. Available: https://doi.org/10.1145/3411764.3445708 12

[39] “Estudos empíricos sobre agentes conversacionais incorporados,” 2006. 12

[40] S. Wang, T. Xu, H. Li, C. Zhang, J. Liang, J. Tang, P. S. Yu, and Q. Wen, “Large language models for education: A survey and outlook,” 2024. [Online]. Available: https://arxiv.org/abs/2403.18105 12

[41] H. Zhao, Z. Liu, Z. Wu, Y. Li, T. Yang, P. Shu, S. Xu, H. Dai, L. Zhao, H. Jiang, Y. Pan, J. Chen, Y. Zhou, G. Mai, N. Liu, and T. Liu, “Revolutionizing finance with llms: An overview of applications and insights,” 2024. [Online]. Available: https://arxiv.org/abs/2401.11641 12

[42] J. Savelka, A. Agarwal, M. An, C. Bogart, and M. Sakr, “Thrilled by your progress! large language models (gpt-4) no longer struggle to pass assessments in higher education programming courses,” in *Proceedings of the 2023 ACM Conference on International Computing Education Research V.1*, ser. ICER 2023. ACM, Aug. 2023. [Online]. Available: http://dx.doi.org/10.1145/3568813.3600142 12

[43] A. Dalal, U. Venaik, D. R. Kumari, and A. Venaik, “Chatgpt’s role in healthcare education, research, and practice: A systematic review of promising prospects and legitimate concerns,” *Educational Administration Theory and Practice journal*, vol. 29, pp. 337–344, 01 2023. 13

[44] Y. Hua, F. Liu, K. Yang, Z. Li, H. Na, Y. han Sheu, P. Zhou, L. V. Moran, S. Ananiadou, A. Beam, and J. Torous, “Large language models in mental health care: a scoping review,” 2024. [Online]. Available: https://arxiv.org/abs/2401.02984 13

[45] J. Zhou, C. Zheng, B. Wang, Z. Zhang, and M. Huang, “Case: Aligning coarse-to-fine cognition and affection for empathetic response generation,” 2023. [Online]. Available: https://arxiv.org/abs/2208.08845 13

[46] H. Qiu, H. He, S. Zhang, A. Li, and Z. Lan, “Smile: Single-turn to multi-turn inclusive language expansion via chatgpt for mental health support,” 2024. [Online]. Available: https://arxiv.org/abs/2305.00450 13

[47] H. Sun, Z. Lin, C. Zheng, S. Liu, and M. Huang, “Psyqa: A chinese dataset for generating long counseling text for mental health support,” 2021. [Online]. Available: https://arxiv.org/abs/2106.01702 13

[48] J. M. Liu, D. Li, H. Cao, T. Ren, Z. Liao, and J. Wu, “Chatcounselor: A large language models for mental health support,” 2023. [Online]. Available: https://arxiv.org/abs/2309.15461 13

[49] H. Na, “Cbt-llm: A chinese large language model for cognitive behavioral therapy-based mental health question answering,” 2024. [Online]. Available: https://arxiv.org/abs/2403.16008 13

[50] S. Lee, S. Kim, M. Kim, D. Kang, D. Yang, H. Kim, M. Kang, D. Jung, M. H. Kim, S. Lee, K.-M. Chung, Y. Yu, D. Lee, and J. Yeo, “Cactus: Towards psychological counseling conversations using cognitive behavioral theory,” 2024. [Online]. Available: https://arxiv.org/abs/2407.03103 13

[51] □□□, □□□, □□□, □□□, and S. Hahn, “Enhancing empathic reasoning of large language models based on psychotherapy models for ai-assisted social support,” □□□□, no. 1, pp. 23 – 48, 2024. [Online]. Available: http://dx.doi.org/10.19066/cogsci.2024.35.1.002 13

[52] P. Guide, “Prompt engineering guide,” 2024, accessed: 2025-01-29. [Online]. Available: https://www.promptingguide.ai/ 14

[53] Amazon Web Services. (2024) What is prompt engineering? [Online]. Available: https://aws.amazon.com/what-is/prompt-engineering/ 14

[54] G. M. Muktadir, “A brief history of prompt: Leveraging language models. (through advanced prompting),” 2023. [Online]. Available: https://arxiv.org/abs/2310.04438 14, 15

[55] A. Bin Sawad, B. Narayan, A. Alnefaie, A. Maqbool, I. Mckie, J. Smith, B. Yuksel, D. Puthal, M. Prasad, and A. B. Kocaballi, “A systematic review on healthcare artificial intelligent conversational agents for chronic conditions,” *Sensors*, vol. 22, no. 7, 2022. [Online]. Available: https://www.mdpi.com/1424-8220/22/7/2625 15

[56] J. Oliveira, T. Silva, R. Oliveira, and E. Furtado, “Recommendations of embodied conversational agents to healthcare applications,” in *Proceedings of the 11th International Conference on Information and Communication Technologies for Ageing Well and e-Health - ICT4AWE*, INSTICC. SciTePress, 2025, pp. 352–359. 15

[57] G. M. Linders, J. Vaitonytundefined, M. Alimardani, K. O. Mitev, and M. M. Louwerse, “A realistic, multimodal virtual agent for the healthcare domain,” in *Proceedings of the 22nd ACM International Conference on Intelligent Virtual Agents*, ser. IVA ’22. New York, NY, USA: Association for Computing Machinery, 2022. [Online]. Available: https://doi.org/10.1145/3514197.3551250 15

[58] M. L. Lupetti, E. Hagens, W. Van Der Maden, R. Steegers-Theunissen, and M. Rousian, “Trustworthy embodied conversational agents for healthcare: A design exploration of embodied conversational agents for the periconception period at erasmus mc,” in *Proceedings of the 5th International Conference on Conversational User Interfaces*, ser. CUI ’23. New York, NY, USA: Association for Computing Machinery, 2023. [Online]. Available: https://doi.org/10.1145/3571884.3597128 15

[59] J. Simpson, F. Gaiser, M. Macík, and T. Breßgott, “Daisy: A friendly conversational agent for older adults,” in *Proceedings of the 2nd Conference on Conversational User Interfaces*, ser. CUI ’20. New York, NY, USA: Association for Computing Machinery, 2020. [Online]. Available: https://doi.org/10.1145/3405755.3406166 15

[60] J. Fruitet, M. Fouillen, V. Facque, H. Chainay, S. De Chalvron, and F. Tarpin-Bernard, “Engaging with an embodied conversational agent in a computerized cognitive training: an acceptability study with the elderly,” in *Companion Publication of the 25th International Conference on Multimodal Interaction*, ser. ICMI ’23 Companion. New York, NY, USA: Association for Computing Machinery, 2023, p. 359–362. [Online]. Available: https://doi.org/10.1145/3610661.3616130 15

[61] P. Sajjadi, L. Hoffmann, P. Cimiano, and S. Kopp, “A personality-based emotional model for embodied conversational agents: Effects on perceived social presence and game experience of users,” *Entertainment Computing*, vol. 32, p. 100313, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1875952119300011 15

[62] H. Gaffney, W. Mansell, and S. Tai, “Conversational agents in the treatment of mental health problems: Mixed-method systematic review,” *JMIR Ment Health*, vol. 6, no. 10, p. e14166, Oct 2019. [Online]. Available: https://mental.jmir.org/2019/10/e14166 16

[63] M. Milne-Ives, C. de Cock, E. Lim, M. H. Shehadeh, N. de Pennington, G. Mole, E. Normando, and E. Meinert, “The effectiveness of artificial intelligence conversational agents in health care: Systematic review,” *J Med Internet Res*, vol. 22, no. 10, p. e20346, Oct 2020. [Online]. Available: http://www.jmir.org/2020/10/e20346/ 16

[64] M. R. Scholten, S. M. Kelders, and J. E. Van Gemert-Pijnen, “Self-guided web-based interventions: Scoping review on user needs and the potential of embodied

conversational agents to address them," *J Med Internet Res*, vol. 19, no. 11, p. e383, Nov 2017. [Online]. Available: http://www.jmir.org/2017/11/e383/ 16

[65] S. Provoost, H. M. Lau, J. Ruwaard, and H. Riper, "Embodied conversational agents in clinical psychology: A scoping review," *J Med Internet Res*, vol. 19, no. 5, p. e151, May 2017. [Online]. Available: http://www.jmir.org/2017/5/e151/ 16

[66] S. Nelekar, A. Abdulrahman, M. Gupta, and D. Richards, "Effectiveness of embodied conversational agents for managing academic stress at an indian university (aru) during covid-19," *British Journal of Educational Technology*, vol. 53, no. 3, pp. 491–511, 2022. [Online]. Available: https://bera-journals.onlinelibrary.wiley.com/doi/abs/10.1111/bjet.13174 16

[67] C. Lisetti, "10 advantages of using avatars in patient-centered computer-based interventions for behavior change," *SIGHIT Rec.*, vol. 2, no. 1, p. 28, Mar. 2012. [Online]. Available: https://doi.org/10.1145/2180796.2180820 16

[68] K. K. Fitzpatrick, A. Darcy, and M. Vierhile, "Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): A randomized controlled trial," *JMIR Ment Health*, vol. 4, no. 2, p. e19, Jun 2017. [Online]. Available: http://mental.jmir.org/2017/2/e19/ 16

[69] J. Martínez-Miranda, A. Martínez, R. Ramos, H. Aguilar, L. Jiménez, H. Arias, G. Rosales, and E. Valencia, "Assessment of users' acceptability of a mobile-based embodied conversational agent for the prevention and detection of suicidal behaviour," *Journal of Medical Systems*, vol. 43, no. 8, p. 246, 2019. [Online]. Available: https://doi.org/10.1007/s10916-019-1387-1 16

[70] Z. Ma, Y. Mei, and Z. Su, "Understanding the benefits and challenges of using large language model-based conversational agents for mental well-being support," 2023. [Online]. Available: https://arxiv.org/abs/2307.15810 16

[71] J. Mercado, I. E. Espinosa-Curiel, and J. Martínez-Miranda, "Embodied conversational agents providing motivational interviewing to improve health-related behaviors: Scoping review," *J Med Internet Res*, vol. 25, p. e52097, Dec 2023. [Online]. Available: https://www.jmir.org/2023/1/e52097 16

[72] J. Benita, S. Jaswanth, N. Bhuvaneshwar, R. Yuvaraj, and Y. L. Narayana, "Phoenix: A conversational agent for emotional well-being and psychological support," in *2025 International Conference on Multi-Agent Systems for Collaborative Intelligence (ICMSCI)*, 2025, pp. 1137–1142. 17

[73] P. M. de Souza, I. d. C. Pires, V. G. Motti, H. M. Caseli, J. B. Neto, L. C. Martini, and V. P. de Almeida Neris, "Design recommendations for chatbots to support people with depression," in *Proceedings of the 21st Brazilian Symposium on Human Factors in*

Computing Systems, ser. IHC '22. New York, NY, USA: Association for Computing Machinery, 2022. [Online]. Available: https://doi.org/10.1145/3554364.3559119 17

[74] W. Vossen, M. Szymanski, and K. Verbert, "The effect of personalizing a psychotherapy conversational agent on therapeutic bond and usage intentions," in *Proceedings of the 29th International Conference on Intelligent User Interfaces*, ser. IUI '24. New York, NY, USA: Association for Computing Machinery, 2024, p. 761–771. [Online]. Available: https://doi.org/10.1145/3640543.3645195 17

[75] A. Melo, I. Silva, and J. Lopes, "Chatgpt: A Pilot Study on a Promising Tool for Mental Health Support in Psychiatric Inpatient Care," *International Journal of Psychiatric Trainees*, vol. 2, no. 2, feb 9 2024. 17

[76] M. Barreda, D. Cantarero-Prieto, D. Coca, A. Delgado, P. Lanza-León, J. Lera, R. Montalbán, and F. Pérez, "Transforming healthcare with chatbots: Uses and applications—a scoping review," *DIGITAL HEALTH*, vol. 11, p. 20552076251319174, 2025. [Online]. Available: https://doi.org/10.1177/20552076251319174 18

[77] J. Weizenbaum, "Eliza—a computer program for the study of natural language communication between man and machine," *Commun. ACM*, vol. 9, no. 1, p. 36–45, Jan. 1966. [Online]. Available: https://doi.org/10.1145/365153.365168 18

[78] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, "Lexicon-based methods for sentiment analysis," *Computational Linguistics*, vol. 37, no. 2, pp. 267–307, 06 2011. [Online]. Available: https://doi.org/10.1162/COLI_a_00049 18

[79] A. Esuli and F. Sebastiani, "Sentiwordnet: A publicly available lexical resource for opinion mining," in *International Conference on Language Resources and Evaluation*, 2006. [Online]. Available: https://api.semanticscholar.org/CorpusID:6247656 18

[80] M. E. Moussa, E. H. Mohamed, and M. H. Haggag, "Opinion mining: a hybrid framework based on lexicon and machine learning approaches," *International Journal of Computers and Applications*, vol. 43, no. 8, pp. 786–794, 2021. [Online]. Available: https://doi.org/10.1080/1206212X.2019.1615250 18

[81] A. J. D. D. and J. M. M., "A novel sentiment analysis for amazon data with tsa based feature selection," *Scalable Computing: Practice and Experience*, vol. 22, no. 1, pp. 53–66, 2021. [Online]. Available: http://www.scpe.org 18

[82] A. Phadte and M. L. Dhore, "Sentmixsent: A comparative study of sentiment analysis techniques for code-mixed text," in *2025 International Conference on Data Science and Business Systems (ICDSBS)*, 2025, pp. 1–11. 18

[83] S. K. Bharti, S. Varadhaganapathy, R. K. Gupta, P. K. Shukla, M. Bouye, S. K. Hingaa, and A. Mahmoud, "Text-based emotion recognition using deep learning approach," *Computational Intelligence and Neuroscience*, vol. 2022, no. 1, p. 2645381, 2022. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1155/2022/2645381 18

[84] R. Hegde, S. Kumar Hegde, Ranjani, and S. S, “Enhancing emotion recognition using advanced speech processing techniques,” in *2025 International Conference on Multi-Agent Systems for Collaborative Intelligence (ICMSCI)*, 2025, pp. 1565–1571. 18

[85] E. Omara, M. Mousa, and N. Ismail, “Character gated recurrent neural networks for arabic sentiment analysis,” *Scientific Reports*, vol. 12, no. 1, p. 9779, 2022. [Online]. Available: https://doi.org/10.1038/s41598-022-13153-w 18

[86] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “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)*, J. Burstein, C. Doran, and T. Solorio, Eds. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423/ 18

[87] C. Lalrinawma and M. V. Nunsanga, “Sentiment analysis review: Methods and challenges,” in *2025 3rd International Conference on Intelligent Systems, Advanced Computing and Communication (ISACC)*, 2025, pp. 831–838. 18

[88] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, “Roberta: A robustly optimized bert pretraining approach,” *ArXiv*, vol. abs/1907.11692, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:198953378 18

[89] *Kashf Journal of Multidisciplinary Research*, vol. 2, no. 07, p. 75–86, Jul. 2025. [Online]. Available: https://kjmr.com.pk/kjmr/article/view/536 18