VIETNAM NATIONAL UNIVERSITY OF HO CHI MINH CITY
THE INTERNATIONAL UNIVERSITY
SCHOOL OF COMPUTER SCIENCE AND ENGINEERING

Chatbot for Answering Mental Health Questions

By
Trịnh Tiến Đạt
ITDSIU20109

A thesis submitted to the School of Computer Science and Engineering in partial fulfillment of the requirements for the degree of Bachelor of Information Technology/Computer Science/Computer Engineering

Ho Chi Minh City, Vietnam
2025

Chatbot for Answering Mental Health Questions

APPROVED BY:

_____

Ho Long Van
(Supervisor)

_____

Tran Thanh Tung
(Committee Chair)

_____

Vi Chi Thanh
(Committee Member)

THESIS COMMITTEE

ACKNOWLEGMENTS

I extend my deepest gratitude to Dr. Van Ho Long, whose unwavering guidance, encouragement, and invaluable knowledge were instrumental throughout the writing process of this graduation thesis. His diligent mentorship, incisive professional comments, and consistent encouragement provided immense motivation, helping me overcome challenges and successfully complete my research.

I would also like to express my sincere appreciation to the faculty members of the School of Computer Science and Engineering, International University – Vietnam National University, Ho Chi Minh City, for their support and for creating a conducive environment during my studies and research.

Finally, my heartfelt thanks go to the Thesis Evaluation Committee for their time, insightful comments, and valuable feedback, which significantly contributed to the improvement of this thesis.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Mental disorders such as depression, anxiety, and stress pose a significant global health challenge, affecting millions worldwide. Despite their increasing prevalence, access to mental healthcare remains limited due to a shortage of trained specialists, high treatment costs, and persistent social stigma. AI-powered chatbots have emerged as a promising avenue for scalable and accessible mental health support; however, current implementations often lack empathy, context sensitivity, and therapeutic efficacy.

This study aims to develop and evaluate a more advanced AI-based chatbot platform to enhance mental healthcare. Our approach utilizes a Retrieval-Augmented Generation (RAG) architecture combined with fine-tuned Large Language Models (LLMs), specifically Qwen-2.5, Llama-3, and Gemma-2. The primary objectives were to design a RAG system featuring an optimized hybrid retrieval module—integrating BM25 keyword search and semantic embedding search, followed by reranking—and to assess the benefits of fine-tuning the selected LLMs on a domain-specific mental health corpus (approximately 15,000 examples) using the QLoRA method. The performance of various retrieval configurations was evaluated using metrics such as Precision, Recall, Mean Average Precision (MAP), and Mean Reciprocal Rank (MRR). The impact of fine-tuning on LLMs was quantified by comparing the performance of pre-trained and fine-tuned models using ROUGE, BLEU, and METEOR scores.

Results demonstrate that the hybrid retrieval system with a reranker significantly outperformed other methods in delivering contextually accurate content, achieving an MAP score of 0.6704 and an MRR of 0.9655. In the LLM evaluation, the fine-tuned Llama-3 model exhibited the superior performance, particularly in semantic relevance, as indicated by a METEOR score of 0.2453. This result suggests that the fine-tuned model has an enhanced ability to generate empathetic and contextually appropriate responses, outperforming both its pre-trained counterpart and the other fine-tuned models. This work clearly illustrates the feasibility and advantages of integrating a state-of-the-art RAG architecture with well-configured LLMs to develop more effective mental health support chatbots. The resulting system and empirical findings offer a promising pathway for creating AI tools capable of providing more accurate, compassionate, and readily accessible mental healthcare, thereby contributing to the global effort against mental health challenges.

CHAPTER 1
INTRODUCTION

1.1. Background

Mental illnesses like anxiety, depression, and stress-related disorders pose a global health problem, with over 280 million people worldwide affected **[1]**. Despite their increasing prevalence, mental healthcare remains elusive due to shortages of skilled personnel, high treatment costs, and ongoing social stigmatization. These barriers have driven the development of innovative and scalable solutions, particularly digital technologies, to deliver timely and accessible mental health support.

In recent years, conversational agents, or chatbots, have emerged as a promising tool for delivering automated, around-the-clock, and stigma-free mental health assistance. Unlike traditional interventions, chatbots can provide immediate responses and personalized interactions, making them especially suitable for individuals seeking support in real time. The use of artificial intelligence (AI) has significantly enhanced their capabilities, enabling them to detect user emotions and give contextually appropriate responses. Core AI technologies such as Natural Language Processing (NLP) and Large Language Models (LLMs) have enabled chatbots to interpret user intent, understand users, and offer evidence-based mental health guidance. Moreover, the application of Retrieval-Augmented Generation (RAG) allows chatbots to look up and incorporate relevant information from external sources, thereby improving the depth and relevance of their responses.

The latest large language model developments, e.g., Qwen-2.5, Llama-3, and Gemma-2, have proven capable of achieving tremendous ability for conversation due to the fact that they are able to generate coherent and natural-sounding responses. However, their applicability in niche domains, such as mental health counseling, will need to be adapted with niche datasets for better accuracy, emotional intelligence, and contextual appropriateness. Evaluation of the performance of these models is also necessary, with metrics such as ROUGE, BLEU, and METEOR offering quantitative assessments of their linguistic quality and contextual appropriateness. Despite such technological breakthroughs, there is limited comparative research on analyzing the performance of different LLMs for mental health applications, particularly for model comparison before and after fine-tuning with material related to mental health.

The present study bridges this gap by investigating the development of chatbots for mental health support based on AI through Retrieval-Augmented Generation (RAG), Natural Language Processing (NLP), and Large Language Models (LLMs). By comparing model performance prior to and subsequent to fine-tuning on mental health-specific data and measuring them against metrics such as ROUGE, BLEU, and METEOR, this study aims to be at the growing intersection of AI and mental health. It is expected that the findings will inform the construction of smarter and more compassionate chatbots, eventually making mental health care more accessible and of higher quality worldwide.

1.2. Problem Statement

Alhough AI chatbots have the immense potential to provide affordable and accessible mental health support, current implementations are unable to meet the complex psychological and emotional needs of users. The core problem is the quality of response from Large Language Models (LLMs) such as Qwen-2.5, Llama-3, and Gemma-2. While these models are able to

talk quite naturally, they lack empathy, context sensitivity, and depth in therapeutic responses, which results in generic responses that lack therapeutic value.

The primary reason for this limitation is the lack of domain-specific fine-tuning. The generic data-trained LLMs fail to capture the nuance of mental health language, i.e., the emotional undertones, and evidence-based interventions. Even the field lacks thorough comparative analyses of the performance of different LLMs in mental health settings, which hinders the selection and optimization of the most suitable models.

The measurement of chatbot performance is also a priority concern. Traditional NLP metrics such as ROUGE and BLEU are lacking when measuring essential facets of therapeutic communication such as empathy and contextual appropriateness. Despite the presence of more recent metrics such as METEOR, the lack of a standard evaluation system has slowed comparative analysis and measuring the true efficacy of chatbots.

1.3. Scope and Objectives

The research aims to develop and validate the effectiveness of AI-driven chatbots for better mental healthcare, addressing three common and high-impact conditions, namely depression, stress, and autism spectrum disorder (ASD). This thesis focuses solely on designing, implementing, and testing the communication capabilities of the AI chatbot to provide initial feedback, recommendations, and information support to individuals experiencing issues relating to the three selected mental health disorders. The study will not focus on exact medical diagnosis or provision of intensive care but on the role played by chatbots as a complementary support mechanism to help ease the burden on the traditional healthcare system and increase access to early support. The training and test data for the chatbot will be accurately selected such that the language and emotional nuances pertaining to depression, stress, and autism are covered and fitting.

The key goal of this research is to develop an AI chatbot system that can have empathetic and suitable interaction with problem users experiencing depression, stress, or autism spectrum disorders. The system will educate them about useful facts, teach basic coping techniques, and suggest additional sources of help. For this, the study will attempt to build and develop the fundamental AI features of the chatbot, from a specialized mental health database, in such a way that the chatbot can comprehend and respond to every provided context correctly and rightfully. In addition, the behavior of the chatbot will also be comprehensively tested using a tailor-made evaluation framework that does not just verify linguistic accuracy but also approximates the empathetic level, suitability for the therapy environment, and ability to provide meaningful support to the users. Based on the outcomes of the research, the team will recommend pragmatic suggestions for enhancing the efficiency in creating and rolling out AI chatbots for mental health care, particularly for the three conditions under examination.

By its focus on depression, stress, and autism spectrum disorders, this study seeks to generate a helpful first-line support tool that can reduce barriers in seeking help and improve the quality of life of sufferers of these mental conditions.

1.4. Structure of thesis

In order to report the thesis research progress in a coherent and organized way, the report falls under main chapters, each of which plays a particular function in the formulation and presentation of the findings. The outline is designed to take the reader through the summary of

the issue to solutions provided, their application, results achieved, and lastly conclusions and recommendations for additional research.

In particular, the thesis structure consists of the following chapters:

- Chapter 1: Introduction: This is an introductory chapter that gives the background of the research issue. This chapter will offer the general background of the use of AI chatbots in mental health care, the significance of this issue in the contemporary world. More specifically, this chapter will state the Problem Statement – with discussion of the current problems and limitations in this area, and establish the Scope and some Objectives which the research is supposed to achieve.

- Chapter 2: Literature Review: This chapter will present the foundation knowledge for the thesis. The chapter presents introductory principles of artificial intelligence, natural language processing (NLP) and more specifically Large Language Models (LLM) – the technology upon which chatbots are based. The chapter will then give a critical review of the existing literature on AI chatbot use in mental health. This will enable the reader to know what has been accomplished, what has been realized and what gaps exist that this research aims to bridge.

- Chapter 3: Methodology: This is the vital chapter, explaining how research was carried out. This chapter will reveal the data gathering and processing procedure for creating a training dataset of mental health. This chapter will also illustrate the selection, tuning, and integration procedure of LLM models within the chatbot system. This chapter will also give a description of the tools, libraries, and technologies used in the development process, and the design of the suggested chatbot system.

- Chapter 4: Experiments and Evaluation: This chapter will emphasize the real performance and experiment evaluation of the chatbot. It will describe the experimental scenarios created to test the chatbot's ability in supporting some mental health issues. In particular, this chapter will describe the evaluation framework used, e.g., the evaluation metrics and criteria (e.g., empathy, contextual relevance, and accuracy) for determining the effectiveness of the generated responses of the chatbot. Findings of experiment and evaluation analysis will be reported objectively.

- Chapter 5: Discussion: This chapter will discuss the implications of the results, the new understanding gleaned from the research, and how it can contribute to the practice of mental health counseling.

- Chapter 6: Conclusion and Future Work: This last chapter will provide the thesis summary of the key contributions. This chapter will also summarize the key findings from the research and review if it has been effective in addressing the problem outlined. Furthermore, this chapter will also suggest possible future research opportunities, paving the way for the ongoing development and innovation of AI chatbots in mental health.

CHAPTER 2

LITERATURE REVIEW

2.1. Mental Health Challenges and Digital Solutions for Access

One of the most significant challenges in addressing the global mental health crisis is the persistent issue of limited access to care. Inflexible service delivery models, a shortage of qualified professionals, and geographical barriers frequently hinder individuals from receiving timely and effective mental health support. Research on mental health service provision in the US public schools **[2]** highlights a stark disparity: rural schools are significantly less likely to offer diagnostic mental health assessments compared to their metropolitan counterparts. This disparity is primarily attributed to limited access to mental health specialists and insufficient funding—factors more commonly reported as major constraints by rural schools. Increasingly, technology-driven solutions are being recognized as promising tools to help overcome these barriers and enhance the accessibility of mental health services.

Getting Past Geographical Barriers: Traditional mental health treatment frequently necessitates inperson visits to clinics or practitioners, which presents a significant and multifaceted challenge for individuals residing in rural or isolated areas where mental health professionals and specialized facilities may be scarce or even nonexistent. The impact of these geographical constraints is evident in research from diverse global contexts. For instance, a study examining the distribution of mental health resources **[3]** in Hunan Province, China, revealed a clear urban-rural disparity, with a concentration of facilities and professionals in more developed cities, leaving rural populations with significantly less access to necessary care. Similarly, research exploring the experiences of emergency department staff in a rural Australian hospital **[4]** highlighted the critical issue of a lack of specialist mental health expertise within these geographically isolated settings, further hindering the provision of timely and appropriate support.

However, the landscape of mental health service delivery has been significantly impacted by the rise of telehealth, particularly accelerated by the recent global pandemic. An international survey of mental health professionals across 100 countries revealed a dramatic increase in the adoption and utilization of telehealth for clinical services since the onset of COVID-19, with over 90% of respondents reporting initiating or increasing their use of remote modalities **[5]**. This widespread adoption underscores the potential of telehealth to overcome geographical limitations by enabling individuals to interact with therapists, counselors, and support groups remotely through video conferencing, phone calls, or text messaging. The convenience and privacy afforded by virtual accessibility allow individuals to receive support in their homes, eliminating the need for arduous travel and reducing the associated time and financial burdens of traditional in-person visits.

Furthermore, the positive perceptions of telehealth among professionals suggest a growing acceptance of its effectiveness and patient satisfaction. While concerns regarding quality of care compared to in-person services and technical issues remain, the rapid shift towards telehealth highlights its viability as a crucial tool in expanding the reach of mental health support, particularly to underserved populations in geographically remote areas. As the findings suggest, continued investment in training and the development of best practices for telehealth is essential to further optimize its potential in addressing the global mental health challenge and ensuring equitable access to care, regardless of location.

Advantages in technology have provided a variety of digital interventions for mental health. AI chatbots have emerged as a promising tool for providing relief through automated, 24/7, and non-stigmatizing care. The utilization of the recent scoping review, "AI Chatbots for Mental Health: A Scoping Review of Effectiveness, Feasibility, and Applications," **[38]** synthesized the evidence for the effectiveness, feasibility, and principal applications areas of AI chatbots for mental health. Similarly, systematic review "Evaluating Generative AI in Mental Health: Systematic Review of Capabilities and Limitations" **[39]** explained the necessary capabilities and limitations of generative AI in mental health, which is assessing safety- and reliability-related issues.

The escalating global mental health crisis is significantly exacerbated by a severe and widely acknowledged shortage of qualified mental health practitioners. World Health Organization data starkly **[6]** illustrate this imbalance, revealing a disproportionately low number of professionals relative to the prevalence of mental health disorders worldwide. To effectively address this critical deficit, innovative technological solutions, particularly those leveraging artificial intelligence, offer considerable promise. As insightfully explored by Sim and Choo (2025) **[7]** in their vision of an AI-enhanced mental health ecosystem, AI, especially through the sophisticated capabilities of large language models, presents a unique opportunity to complement traditional human-led interventions by providing scalable and contextually intelligent support. AI-powered chatbots, for example, can function as an initial point of engagement, delivering psychoeducation, fundamental coping mechanisms, and facilitating self-monitoring, thereby strategically alleviating the workload on human clinicians and allowing them to focus on more complex and urgent cases. The concept of AI as a virtual peer, counselor, or therapist, as highlighted by Sim and Choo, further underscores this potential, with existing AI chatbots demonstrating efficacy in reducing symptoms of anxiety and depression. Moreover, by transcending geographical constraints **[3]**, **[4]**, internet-based therapy platforms can connect individuals with a more extensive network of therapists, effectively mitigating the challenges posed by the uneven distribution of mental health professionals. Ultimately, the intelligent and responsible integration of AI-driven tools holds the key to optimizing the efficiency of our limited mental health workforce and ensuring more widespread access to crucial support.

Improving Convenience and Flexibility: Conventional mental health services often follow rigid schedules, requiring individuals to take time off work or rearrange their daily calendars to attend appointments. In contrast, digital mental health interventions (DMHIs) offer increased convenience and flexibility, allowing users to access support whenever and wherever it best fits their needs **[8]**.

A review by Vaidyam et al. (2019) **[9]** highlights the growing role of artificial intelligence (AI), particularly intelligent chatbots, in enhancing accessibility to mental health services. These AIpowered chatbots, such as Woebot or Anna from Happify Health, can provide on-demand access to mood monitoring, crisis assistance, and self-help resources—empowering users to take charge of their mental well-being and seek timely support. In addition to accommodating busy lifestyles, asynchronous communication methods, such as encrypted messaging with therapists or chatbotguided sessions, enable continuous and personalized support between scheduled appointments.

The review emphasizes that chatbots in DMHIs have shown promise in promoting user engagement, increasing accountability, and potentially improving mental health outcomes, particularly in managing symptoms of depression and anxiety. However, the authors also stress the importance of addressing ethical, contextual, and design-related concerns—such as natural

language misinterpretation, user privacy, and the limitations of empathetic response—to fully realize the potential of AI in mental health care.

2.2. Overcoming Stigma through Anonymous Digital Support

Mental health stigma is a major barrier to seeking and maintaining treatment for mental and substance use disorders according to Ahmedani, 2025 **[9]**. Goffman described stigma as a "deeply discrediting attribute," while Dudley defined stigma as negative attitudes toward individuals or groups due to characteristics that are perceived as different or inferior to social norms. Despite the availability of effective treatments, stigma often prevents people from getting the help they need.

Forms of Stigma:

- Peril: People with mental illness are often viewed as dangerous or unpredictable.
- Origin: Misbeliefs persist that mental illness is purely under personal control, despite biological evidence.
- Controllability: Individuals are unfairly blamed for not overcoming their condition on their own.
- Concealability: Disorders with visible symptoms tend to face more stigma.
- Course & Stability: Doubts about recovery or treatment success contribute to stigma.
- Disruptiveness: The perceived social disruption caused by mental illness increases stigma.
- Aesthetics: Some symptoms are seen as unpleasant, leading to further bias.

Negative Impacts:

- Avoiding help due to fear of judgment or discrimination.
- Delayed diagnosis and intervention, harming treatment outcomes.
- Limited opportunities in education, employment, and relationships.
- Increased loneliness and risk of suicide due to self-stigma.
- Hindered recovery from reduced self-esteem and social support.
- Reducing stigma is a key role of social work, requiring action through practice, education, policy, and research. Understanding stigma's forms and effects is essential for developing effective interventions.

As stigma continues to be a major barrier to getting mental health care, digital mental health interventions (DMHIs) have emerged as a potential approach to minimize the detrimental effects of stigma. One of DMHIs' primary advantages is their capacity to give users with an anonymous or semi-anonymous environment **[10]**. This anonymity not only minimizes the risk of being evaluated or discriminated against, but it also encourages users to seek information

and support without fear of disclosing their identify, which is especially important in sensitive contexts like work.

Carolan and de Visser's study of employees' perspectives on DMHI **[10]** in the workplace found that anonymity was a particular perspective described by participants. DMHI allows users to engage and receive follow-up support without revealing their true identity to colleagues or sometimes even service providers.

Reduced anxiety about judgment: Anonymity plays an important role in reducing anxiety about judgment and any perception of mental health issues, especially in sensitive environments such as the workplace. Participants expressed concerns about letting colleagues know that they were struggling, fearing being seen as incompetent to cope, and impacting their chances of promotion. DMHIs, with their hidden computers, help users feel safer in seeking help without fear of being judged negatively or revealing personal information.

Increased comfort in seeking information and support: The ability to interact anonymously helps users feel more comfortable in initiating a search for mental health information and support resources. Not having to make a phone call or meet in person to make an initial appointment is considered a convenience factor, making DMHIs more accessible. Some people find it easier to "do it" to help themselves through a digital platform without having to talk to someone in person about personal issues. Anonymity can also increase confidence in initiating a consultation, especially for those who find it difficult to share their concerns through face-to-face communication.

The researcher from Strand et al **[11]** suggest that anonymity plays a foundational role in fostering openness and trust during the initial stages of digital mental health engagement. By enabling users to control their level of identity disclosure, anonymous platforms reduce perceived stigma and encourage sharing of sensitive personal experiences.

To further explore this dynamic, Renn et al **[12]** conducted a mixed-methods study examining treatment preferences among 164 adults in the U.S. who had either received or considered psychotherapy for depression. This research offered a nuanced comparison between traditional face-to-face therapy and various forms of digital psychotherapy, including self-guided, peersupported, and clinician-guided online options.

While digital interventions held promise in overcoming access barriers—including cost, geographic limitations, and especially stigma—participants also expressed considerable concerns regarding data privacy and the effectiveness of non-traditional formats. Interestingly, despite acknowledging that anonymity in digital therapy might reduce stigma, many participants still preferred in-person therapy, valuing the direct interpersonal connection and perceiving it as more secure and effective.

Moreover, participants identified stigma (26.2%) as a significant barrier to traditional therapy, yet the potential solution offered by anonymity in digital contexts was not universally embraced due to fears of surveillance, data breaches, and lack of privacy in shared home environments.

Together, these studies illustrate the dual nature of anonymity in digital mental health support: it can lower the psychological threshold for help-seeking and sharing, as shown in Strand et al.'s study **[11]**, but also generate new forms of vulnerability, as highlighted by Renn et al **[12]**. Thus, anonymity is not a universally positive feature but rather a complex factor whose effectiveness depends on users' trust in the platform's security, perceived therapeutic value, and personal preferences for connection.

Building upon the understanding of anonymity's complex impact, several types of digital interventions have emerged that strategically utilize this feature to enhance mental health support.

- One effective type is virtual assistants and chatbots. These tools offer an anonymous initial point of access for mental health support and information, providing a secure space for individuals hesitant to seek in-person help due to fear of stigmatization. Hungerbuehler et al.'s **[13]** pilot study of the Viki chatbot for employee mental health screening in Brazil demonstrated this. The anonymity provided was a key factor in the high response rate (64.2%), encouraging workers to disclose mental health issues they might otherwise conceal. Beyond anonymous data collection, Viki also offered initial support, information, personalized feedback, and referrals, highlighting the potential of chatbots to lower stigma barriers and facilitate help-seeking.

- Another prominent example of anonymous digital support can be found in online forums and communities, such as Reddit. De Choudhury and De's **[14]** study explored mental health discussions on this platform, where temporary 'throwaway' accounts enable users to discuss sensitive topics openly. The research revealed Reddit as a significant platform for diverse emotional, informational, and practical support. Notably, anonymous users, despite potentially negative content, received substantial support, indicating that anonymity can foster open sharing and community engagement without the constraints of personal identity. This highlights the unique role of platforms like Reddit in bridging traditional forums and social networks for mental health support.

While anonymity offers many potential benefits in overcoming stigma barriers and increasing access to digital mental health interventions, its implementation also comes with important considerations and limitations that need to be carefully considered.

Accountability and safety issues: One of the key challenges associated with anonymous environments is ensuring accountability and safety, particularly in emergency situations or when users are at risk of self-harm **[15]**. Research by Ray B Jones and Emily J Ashurst has shown that in anonymous discussion forums, there is a risk of posts that are uninhibited, irresponsible, and unhelpful. Furthermore, in the context of anonymous mobile mental health apps, **[16]** highlights that many apps are ill-equipped to support users in crisis situations such as suicide risk. The lack of effective emergency intervention mechanisms in anonymous environments can delay or prevent users from receiving timely help when they need it most. This raises questions about the responsibility of service providers and platforms in ensuring the safety of anonymous users.

Ability to build trusting relationships: Anonymity can also affect the ability to build deep trusting relationships between users and service providers (if any). While some users may feel more comfortable sharing personal information in an initially anonymous environment, the lack of faceto-face interaction and identification may hinder the development of trust and meaningful connection over time. Research by Carolan and de Visser (2024) **[12]** also suggests that anonymity may make it easier for users to drop out of an intervention due to the lack of face-to-face interaction and accountability. Ray B Jones and Emily J Ashurst noted mixed views on rapport in webcast group therapy, with some concerned that the online "distance" and anonymity may limit connection. Trust building is often based on transparency and personal interaction, which may be undermined in a completely anonymous environment.

Limitations in the Workplace: Research by Carolan and de Visser (2024) **[12]** also points to another limitation of anonymity in the workplace context. In many office environments, especially open offices, maintaining complete anonymity online can be difficult if colleagues can see each other's screens. This can reduce the benefits of anonymity and raise practical privacy concerns. Therefore, balancing the benefits of anonymity in reducing stigma and increasing accessibility with the need for interaction and support is an important factor to consider when designing and developing effective digital mental health interventions in the workplace.

The need for clear privacy information: While anonymity focuses on concealing a user's identity from others, ensuring the privacy and security of personal data that users share on the platform is of utmost importance. The research of Koh J et al, highlighted major concerns about security vulnerabilities and the lack of clear privacy policies in mobile mental health applications. Even when users interact anonymously, they may still share sensitive information, and the lack of security guarantees can lead to the risk of information leakage to third parties. Therefore, service providers need to provide transparent and clear information about how user data is collected, used, and secured in anonymous platforms to build and maintain user trust. The research of Ray B Jones and Emily J Ashurst also emphasized the need to carefully explain who knows about users' identities when proposing new online services **[16]**.

In summary, while anonymity offers significant benefits in accessing mental health support, issues of accountability, safety, the ability to build trusting relationships, and privacy protection need to be carefully addressed to ensure that anonymous digital interventions are truly beneficial and do not pose unintended risks to users.

2.3. The importance of empathy and trust in mental health support

Empathy and trust are the building blocks of effective psychological support interactions, especially in online settings where many people go to find help. Sharma et al. (2020) **[17]** proposed the EPITOME theoretical framework that outlines that empathy can be expressed through three distinct communication mechanisms: Emotional Response (which is marked by warmth and exploration), Interpretation (which is a reflection of emotional understanding), and Exploration (which examines unspoken experiences). Each of the three plays an important role in establishing connection, building trust, and increasing therapeutic effectiveness by making the seeker feel heard and understood.

In addition, empirical evidence illustrates that empathetic engagement is associated with the establishment of more stable relationships between the seeker and the supporter, as shown in instances where the host user follows the supporter after receiving positive stories. A data analysis from the TalkLife platform lends credence to the fundamental role of empathy in online support contexts. These observations point to the necessity of incorporating genuine empathy into mental health support frameworks, including interventions involving artificial intelligence.

Similarly, the article "Building trust in artificial intelligence and new technologies in mental health" **[18]** highlights that, as much as new technologies and artificial intelligence have enormous potential to improve mental healthcare, their worthiness is ultimately reliant on the maintenance and reinforcement of interpersonal relationships, more so the trust aspect between patients and health professionals. Essential elements of such relationships include empathy, unambiguous communication, and clinician availability.

The "AI chasm" **[18]** framework presented in the article illustrates the necessity of paying attention to not only the technical but also the "soft," "qualitative" aspects of interaction in

artificial intelligence, such as empathy and trust. In overcoming this chasm, it is crucial that researchers and developers give particular attention to the technology design and deployment practices that will create, rather than undermine, interpersonal relationships and trust in the practice of mental healthcare.

Doctors play a significant role in bridging the gap between technology and patients. Both the technological innovation and the users must be assessed for trustworthiness. The integration of artificial intelligence and technological innovation into mental health, therefore, must be approached cautiously with a perspective that emphasizes the human factor to ensure that trust is not only preserved but also promoted and fostered. The effectiveness of technological interventions in this area will depend on the capacity to establish and sustain trust among patients and users.

In their exploration of empathy and trust in mental health support, with specific reference to the incorporation of artificial intelligence (AI) technologies, Shen et al. (2023) **[19]** highlight how essential these factors are in mental healthcare. One of the key issues raised in the study relates to the ability of AI systems to present empathetic responses. Particularly, the research tested participants' reactions to AI-created versus human-created stories, and tested whether revealing the origin of the story (AI or human) influences the degree of empathy displayed by listeners.

Results showed that people will express more empathy towards human-generated stories than those produced using AI. This is a major challenge to the use of artificial intelligence in mental health support services because empathy is an essential foundation for building connection and trust. Telling someone that a story is generated by AI can cause users to deliberately alter their degrees of empathy; however, it also reduces the authentic empathy they feel.

This indicates that merely alerting users to their interaction with AI is not sufficient to provide the emotional gap in the interaction.

One of the underlying reasons found in the study relates to artificial intelligence's nature. Large language models, including ChatGPT, lack first-hand experience and instead generate content by analyzing and mimicking linguistic patterns from human-created data. As a result of this lack of first-hand experience, there is a sense of inauthenticity perceived in the AI's output, which, in turn, negatively impacts its ability to generate empathy among users. This implies that while artificial intelligence holds tremendous potential to offer broad-based mental health support, bridging the empathy gap between humans and AI is not solely a function of technological advancement but also of a profound comprehension of human attitudes and interactions with artificial systems.

Researchers and developers need to focus on developing chatbots and AI programs that can demonstrate understanding and empathy on their own, and appreciate the importance of openness in communications to develop and sustain user trust. As Shen et al. argue, it is not sufficient to generate coherent text only; AI systems need to be designed in a manner that enables them to be able to "understand" and respond correctly to human emotional states in a meaningful manner. Here, the study by Cheng et al. (2023) **[20]** highlights the importance of establishing trust in the collaboration between AI chatbots and users. At the beginning, chatbots had negative ratings for being reliable, owing to the lack of transparency, limited hardware capabilities, and a poor user interaction experience.

However, with improvements implemented in the system based on user feedback, chatbots became more user-centric, easier to use, and better at interpreting user needs, leading to an increased rate of positive reviews.

Based on these studies, pragmatic suggestions can be made regarding creating AI chatbots in the field of mental health, namely:

- Displaying "intelligence": Comprehending and replying responsively to intricate emotional states and user problems, without falling back on formulaic answers.

- Creating a "friendly" environment: Easy interface, shared language, and showing concern and empathy. • Ensuring "security": Promising to protect users' personal information to build greater trust.

- Enabling "comfort" and continual improvement: Listening to feedback to continually improve the quality of interactions.

In alignment with this perspective, the most recent research by Katoch et al. (2025) **[21]** titled "From Algorithms to Empathy: Navigating Ethics, Efficacy and User Trust" assesses the use of AIeMHCs in therapy environments.

This work emphasizes the changing role of chatbots to maximize access to mental health services via the delivery of scalable, customized, and effective interventions. However, the viability of such systems is heavily reliant on their ability to simulate empathetic conversations like those between human beings and sustain high ethical standards.

Ethical challenges, such as privacy, transparency issues, and the existence of algorithmic bias, are raised as key stumbling blocks to trust. The research indicates that these can be overcome by having rigorous human oversight, prioritizing data protection, and developing culturally sensitive natural language processing systems. Notably, the study notes that while advanced language models are increasingly able to mimic empathetic dialogue, their inability to experience emotions genuinely constitutes a foundational limitation compared to human therapists. The notion of user trust is explored in-depth from both technical and human-centered perspectives. Technological attributes like accuracy, reliability, consistency, and useful features are important, yet individual factors like ease of use, perceived usefulness, and congruence with sociocultural norms are also essential. This study shows that building trust requires the application of ethical AI principles, human oversight, open communication about limitations, and realistic user expectation management. In short, to be usable and useful, AI-powered mental health technologies must demonstrate their technical proficiency, but must also gain the trust of people through empathic, transparent, and morally-compliant interaction.

2.4. AI Mental Health Tools: Efficacy and Safety Evaluation

❖ Clinical Efficacy

The use of artificial intelligence (AI) in mental healthcare is a paradigm shift in the provision of psychological treatment and support. These technologies can potentially improve access, facilitate early intervention, and personalize care at scale. They range from sophisticated models that can predict mental health risk to AI-driven chatbots that offer instant emotional support. Since AI systems are increasingly being used to augment or even replace traditional mental health treatments, it is a consideration that it is worthily evaluated for its clinical efficacy and safety. This section critically examines existing methodologies for assessing AI within mental health settings, highlights salient concerns such as bias, trust, and regulatory loopholes, and charts the way forward to ensure these advancements are both ethically and ethically responsible.

Clinical effectiveness pertains to the capacity of AI tools to yield positive mental health outcomes, such as a reduction in the severity of depressive and anxiety symptoms, or an enhancement in overall quality of life. Common evaluation methodologies encompass randomized controlled trials (RCTs) and systematic reviews. RCTs compare the effectiveness of AI interventions against standard care protocols, while systematic reviews synthesize data from multiple studies to derive overarching conclusions. Standardized psychometric instruments, such as The Patient Health Questionnaire-9 (PHQ-9) to assess depression and the Generalized Anxiety Disorder-7 (GAD-7) scale for anxiety disorders are commonly used to quantify outcomes.

A recent study published in NEJM AI in 2025 **[22]** conducted an RCT on 106 patients with a major depressive disorder, generalized anxiety disorder, or eating disorders. The findings reported that a generative AI chatbot reduced depressive symptomatology by 51% from a standard care control, as measured using the PHQ-9 scale. This research points out the potential of AI to offer effective interventions, particularly in those regions where mental health professionals are lacking.

A 2025 systematic review of 15 studies and published in BMC Psychiatry **[23]** found that AI interventions, such as predictive models and chatbots, increase patient engagement and allow for tailoring of interventions. One group of users, for example, saw a 67.7% reduction in depressive symptoms after using the Wysa app. The review did comment that some of the trials were hampered by small numbers of patients, restricted longitudinal data, and the use of discordant outcome measures, which made comparison between the results difficult.

A 2024 review article published in PMC **[24]** focusing on AI in positive mental health similarly emphasized that while AI can improve symptom management and patient engagement, studies frequently lack robust validation, suggesting a need for larger and more heterogeneous trials to ascertain the true effectiveness of AI interventions.

Notwithstanding these promising results, current research encounters several limitations. Many trials predominantly focus on prevalent disorders such as depression and anxiety, thereby excluding more complex conditions such as schizophrenia. Furthermore, inconsistencies in outcome measures and short follow-up durations diminish the generalizability of findings. A study published in the Journal of Medical Internet Research (JMIR) Mental Health in 2023 **[25]** revealed that 56 out of 153 AI studies did not report comprehensive details regarding data preprocessing, thereby compromising the reliability of the reported results.

❖ Safety Concerns

Privacy and Data Security: Robust legal regulations such as the General Data Protection Regulation (GDPR) in the EU or the Health Insurance Portability and Accountability Act (HIPAA) in the US govern AI applications to protect user privacy due to the sensitive nature of mental health data. However, several AI applications hide their data handling practices, which raises significant privacy concerns, according to a 2023 World Health Organization (WHO) report **[25]**. For example, if chatbot conversation data is not adequately encrypted, personal information may be exposed.

Algorithmic Bias: Algorithmic bias is a risk associated with artificial intelligence models that are created using biased data sets. A 2024 review article in PMC **[24]** pointed out that when training data primarily come from one cultural or demographic background, AI may make incorrect diagnoses or interventions for patients from other groups, leading to differences in

treatment. For instance, a model that was primarily trained on data from Western populations might not recognize depressive symptoms as they manifest in Asian communities.

Risk of Harm: There is a real concern about the likelihood that AI will provide unsuitable or damaging results. According to a 2024 article by Built In, some AI chatbots have been accused of inciting violent or suicidal behavior among their adolescent users **[26]**. These incidents demonstrate the necessity of stringent safety protocols and human supervision. A 2023 JMIR Mental Health **[25]** study also raised concerns regarding the reliability of AI tools, pointing out that only 21 out of 153 studies included external validation.

❖ Regulatory and Ethical Framework

The possible complexity and dangers in using AI in mental healthcare demand robust regulatory frameworks and ethical guidelines. WHO in 2023 **[27]** advocated for more rigorous evaluation prior to the widespread application of AI tools, demanding practical feasibility and ethical concerns. The American Psychological Association (APA) **[28]** also provides recommendations to psychologists for the ethical application of AI, with proper attention to maintaining professional standards and patient welfare.

Standardization of the methods of assessment is extremely important. Methodological quality of AI studies can be assessed using tools such as the Cochrane Risk of Bias Tool and the Prediction model Risk Of Bias ASsessment Tool (PROBAST) to ensure that safety and efficacy are addressed adequately. A 2024 article in ScienceDirect **[29]** suggests that regulatory bodies give clear guidelines on AI development and deployment in mental health, with mandatory demands of algorithmic transparency and comprehensive risk assessment.

Overall, while AI technologies for mental healthcare present promising opportunities for enhancing access to and effectiveness of treatment, their successful implementation hinges on addressing the significant challenges relating to clinical efficacy and safety. Though studies such as the RCT in NEJM AI **[22]** and the review in BMC Psychiatry **[24]** are promising, concerns such as data privacy, algorithmic bias, and the absence of long-term data need to be considered cautiously. Standardized assessment protocols, increased transparency, and interdisciplinary collaboration enable the effective use of AI to improve mental healthcare while ensuring safety, equity, and accountability.

2.5. The Potential of LLM and RAG in Addressing Existing Challenges

❖ Large Language Models (LLMs)

In the era of fast growth of Artificial Intelligence, Large Language Models (LLMs) have attracted most of the research and applications. For an improved understanding of the strengths, weaknesses, and possibilities of these models, numerous reviews and analyses have been done to present an overall picture of the field. LLMs, exemplified by models such as GPT-4 and Alpaca, possess a very sophisticated capacity to comprehend and generate human-like textual data.

Overview of Large Language Models (LLM):

- For instance, the Sokada (2025) **[36]** review article provided an account of the architectural progress and training approaches of LLM models. It points to the trend for increasing model sizes and the evolution of effective fine-tuning methods, along with illustrating key challenges such as interpretability and bias during training. Similarly,

the Empler AI (2025) **[37]** review focuses on investigating the extensive applications of LLMs in different domains. The article evaluates the efficiency and portability of LLMs in customer service automation, text data analysis, and predictive models. Empler AI discusses the necessity for addressing issues related to AI ethics and data privacy while implementing LLMs in large-scale applications. Reviewers such as Sokada **[36]** and Empler AI **[37]** are helpful in informing the landscape of LLM research and provide the context for exploring more specialized areas, such as the mental health arena that this research is focusing on. But only by intensive empirical research such as ours can one gain a complete insight into the performance of LLMs in a particular situation, such as mental health support, after fine-tuning them with domain-specific data, since such reviews penetrate deeper into the special nuances and demands specific to a narrow application domain.

- Facilitating Access to Mental Healthcare Services: LLMs offer the potential to deliver scalable, affordable, and accessible mental health care solutions through digital means. Scholarly studies on Mental-LLM **[31]** have demonstrated the capability of LLMs to predict mental state disorders based on the analysis of online text information, such as social media dialogue, in an effort to enable early detection and timely intervention even in the absence of providing conventional services (Mental-LLM: Leveraging Large Language Models). These technologies may be easily implemented on ubiquitous personal devices like smartphones and computers, allowing users anywhere in the world round-the-clock access to psychological support, thus alleviating financial costs and limitations caused by geographical distance.

- Mitigating Stigma Through Anonymous Digital Support Modalities: The pervasive effects of social stigma are a deep-seated obstacle that prevents a lot of individuals from receiving necessary mental health treatment. The provision of anonymous internet support in the guise of AI interventions can have a mitigating effect on this endemic problem as users often indicate that they are more comfortable supplying personal information to a neutral technological system. LLMs, due to their inherent capability of generating compassionate and supportive textual responses, can be potential drivers in creating a safe and secure space for users. Research associated with Opportunities and Risks of LLMs in Mental Health **[32]** identifies the capacity of AI-based applications to encourage individuals to seek help by way of non-judgmental communication. For example, chatbots based on LLM can be programmed to employ encouraging words and phrases, such as "I am here to listen" or "You are not alone," in a bid to establish a sense of shared understanding among users. But this literature also serves a caution that when responses from an AI are contradictory or otherwise inappropriate, they can in turn promote an increase in loneliness or foster suspicion.

- Enhancing Empathy and Trust in Artificial Intelligence Interactions: Building trust and empathy in AI interactions is a major psychological and technological challenge. Nonetheless, significant progress is being made in this crucial area thanks to recent developments in LLMs. Modern LLMs are better able to understand context and produce natural language responses that closely mimic human speech patterns. However, since AI systems are incapable of subjective emotional experience, achieving true empathy is still a challenge. According to a thorough analysis of LLMs in mental health care, users may still feel that there is a lack of a strong emotional bond even though these models are capable of accurately simulating sympathetic reactions (LLM Review). In order to overcome this constraint, developers must keep improving LLMs' emotional recognition skills and further incorporate features with particular user data,

like thorough chat histories and psychological profiles, to produce responses that are more sensitive and individualized.

- Ensuring Clinical Efficacy and Safety in AI-Driven Mental Health Interventions: To guarantee that AI solutions used in mental healthcare settings are applicable with real-world benefits and do not unintentionally cause harm, safety and efficacy are medical necessities of utmost importance. Despite their amazing generative powers, LLMs are far from perfect and may produce responses that are factually incorrect or stale. Additionally, research on the use of modified LLM to diagnose mental disorders with clinical notes as input indicates that LLMs can significantly improve diagnostic performance, especially for common conditions like depression and anxiety.

- Limitations and Future Developmental Directions: Despite the important promise realized by LLMs in psychiatric care, several of the current limitations deserve careful scrutiny and concerted research efforts. Firstly, current academic studies lack strong longitudinal evidence concerning the long-term efficacy and safety of these tools when used in real-world clinical practice. A review on arXiv **[34]** emphasizes the necessity of large-scale, prospective clinical trials to rigorously evaluate the long-term clinical impact of LLMs. Secondly, the salient issues of data privacy and algorithmic bias remain fundamental. For instance, if the training data sets that are applied in LLMs are not sufficiently representative of different ethnic or cultural groups, such technology instruments can inadvertently generate inapplicable or prejudiced responses, potentially exacerbating existing health disparities. One such study that was published in PMC **[32]** highlights the necessity of formulating and implementing effective strategies against such inherent threats, such as employing more diverse and representative training data and performing routine model testing and validation practices. Another challenge when fine-tuning LLMs is “catastrophic forgetting,” where the model may forget previously learned information after being trained on new data.

❖ Retrieval-Augmented Generation (RAG)

- Retrieval-Augmented Generation (RAG) is a model that effectively combines information retrieval from outside knowledge bases with text generation. By doing so, RAG generates more accurate and contextually relevant responses. The MDPI survey "Retrieval-Augmented Generation (RAG) Chatbots for Education: A Survey of Applications" **[40]** demonstrated the general usability of RAG across different fields, apart from healthcare.

- Enabling access to mental health services: RAG enables effective, scalable, and affordable mental health care solutions via digital avenues. The "OnRL-RAG: Real-Time Personalized Mental Health Dialogue System" **[35]** and SouLLMate system **[30]** are classic examples, employing RAG to build personalized dialogue support systems, such as suicide risk screening and proactive counseling.

- Lowering Stigma Via Anonymous Digital Support: Use of RAG enhances the trustworthiness of AI interactions by providing answers formulated from verifiable information from permitted sources such as clinical guidelines and mental health education materials. This encourages users to request aid in the initial place.

- Enhancing Empathy and Trust in AI Interactions: RAG enhances the empathy and trust of the system by retrieving suitable data to generate context-specific responses. RAG's provision of specific references or information bases from trusted databases increases

the transparency of AI. The SouLLMate system [30], which employs RAG to build an individualized psychological dialogue system, is a bright example, and in that way, considerably expands user trust and engagement. In order to bring more severe empathy, RAG feature requirements need to be combined with some user information such as chat history and psychological profiles.

- Ensuring clinical effectiveness and safety: In order to safeguard that AI-powered solutions in mental health therapy are doing well and not ill, first and foremost, effectiveness and safety. RAG sidesteps the failures of LLMs by integrating them with extrinsic sources of reliable information such as clinical guidelines and academic literature, significantly improving response consistency and accuracy. An article in NEJM AI **[32]** demonstrates how RAGs have the potential to facilitate improved communication and decision-making in healthcare settings by providing LLMs with access to up-to-date data stores. Lokesh Boggavarapu et al. **[32]**, in their 2024 work, showed that RAG performs well in detecting mental illness using social media words even without provided training datasets. But there must be strict safety guidelines for RAG, such as re-directing users to specialists in the event of high-risk situations being detected.

- Limitations and future directions: Despite the promise of RAG, there are a few limitations. If knowledge bases upon which RAG is built do not fully represent ethnic or cultural groups, then they may give inappropriate or biased answers, which will exacerbate health inequities. A study by PMC **[32]** highlights the need for countermeasures against such dangers, such as regular model testing and validation. RAG can be an active replacement for "catastrophic forgetting" in LLM as it does not modify the base model weights but instead recovers and uses new knowledge from the dynamic KB.

Based on the data in the report, the following table offers a brief synopsis of the mental health chatbot/AI systems discussed in this chapter, including their methodologies, target markets, and main results:

System	Approach	Target User Audience	Main Results
Woebot	AI Chatbot (mentioning the ability to provide ondemand access)	Users searching for mental health support	Provides on-demand access to mood tracking, crisis support, self-help resources. Shows promise in boosting user engagement and potentially improving mental health outcomes
Anna (Happify Health)	Chatbot AI	Users searching for mental health support	Provide on-demand access to mood tracking, crisis support, self-help resources
Viki	Chatbot (chatbot based for screening)	Staff in Brazil	High response rate (64.2%) in employee mental health screening. Provide initial support, information and referral
SouLLMate	Adaptive LLM, Prompt Engineering, RAG	Users searching for mental health support	Real-time personalized mental health dialogue system. Provides a variety of services including suicide risk screening and proactive counseling. Increases user trust and engagement
Mental-LLM	LLM	Analysis of online text data (e.g. social media)	Ability to predict mental state disorders based on online text data analysis

Table 1. Overview of typical AI systems in mental health support

Table 1 indicates an overview of top AI systems that are being used in the mental health support domain. Woebot and Anna (Happify Health), for example, focus on providing on-demand support to clients through mood tracking tools, self-help materials, and crisis support. Viki, however, is used in organizations to screen employees for mental illness with high rates of response. Specifically, SouLLMate employs RAG modeling and Prompt Engineering approaches to produce real-time personalized suggestions, improving user trust and engagement. Finally, Mental-LLM is focused on the ability of online text data analysis in forecasting mental illness, showing great potential for community mental health monitoring. They depict the variety of approaches and applications of AI in mental health care, and the pivotal role of LLM and RAG models in individualizing and improving the quality of care.

CHAPTER 3
METHODOLOGY

3.1. Overview

This chapter presents the general methodology designed and carried out to fulfill the main research goal: designing and properly assessing a chatbot system with current Artificial Intelligence (AI) in order to improve the effectiveness of mental health care. In order to overcome the core challenges and issues that were extensively discussed in Chapters 1 and 2 – especially the urgent necessity for cost-effective, empathetic, and reliable mental health solutions – this study employs a systematic, controlled, and multi-staged procedure: analysis, design, testing, and evaluation.

The chapter begins with an immersion in User Requirements Analysis (Section 3.2). This important first step goes beyond technical requirements to gain a deep familiarity with the needs, expectations, and psychological aspects of potential users. This is a solid foundation upon which to build a chatbot not only functional, but indeed useful and trustworthy.

Based on the discovered requirements, Section 3.3 outlines the System Architecture focused on the proposed Retrieval-Augmented Generation (RAG) model. Within this section, the architecture is explained that unites external knowledge retrieval capabilities with large language models (LLMs) language generation capabilities to output informative and contextually relevant responses.

Lastly, Section 3.4 outlines the System Evaluation Metrics. This section provides an overview of the quantitative and qualitative indicators adopted in the evaluation of chatbot performance. Specific metrics applied to assess the effectiveness of the information retrieval module, the quality of generated responses, and the system's capacity to address key aspects of psychological care including accuracy, reliability, and empathy are comprehensively presented. The selection of such measures was undertaken with caution to make the assessment inclusive, with full awareness that there are challenges in quantifying subjective dimensions of therapeutic communication.

3.2. User Requirement Analysis

An in-depth comprehension of the needs and expectations of the users is essential to the success and helpfulness of a mental health chatbot. Since users approach it for sensitive and varied issues, the analysis of requirements must encompass not just technological functionality but also psychological and emotional factors.

The aim of this phase is to establish a clear set of functional and non-functional requirements, and overall interaction principles, informing the design and development of a chatbot that is able to establish trust, comfort, and deliver substantive supportive value.

In order to achieve this goal, the study will conduct user requirements analysis through the following methods:

- In-depth Literature Review: Following on from the general understanding that has been rigorously synthesized and critiqued in Chapter 2, this stage will synthesize and organize the conclusions of prior research associated with:

- Specific user expectations for digital mental health supportive tools are to be heard, to be understood, to be given accurate information, and to be offered simple coping strategies.
- Barriers and facilitators to the uptake and utilization of chatbots (e.g., stigma, need for anonymity, privacy concerns, need for an engaging interface and instant feedback).
- The significance of empathy, trust, and interpersonal interaction factors in forging successful therapeutic relationships between chatbots and users.
- The need for seamless, on-demand accessibility, particularly by some groups who encounter barriers in accessing mainstream services.

- Identification of Target User Groups and Their Specific Needs: Although the chatbot is designed to assist a wide range of general psychological problems, the study will aim to assist user groups that are typically challenged by issues, for example, mild to moderate stress and anxiety, and those who require initial mental health information and guidance. This identification simplifies the process of outlining interaction scenarios and the nature of assistance the chatbot should offer.

From the analysis above, a comprehensive list of requirements will be generated, including:

- Functional Requirements:
  - Provide reliable and contextual information.
  - Create feedback with empathetic and appropriate language.
  - Maintain the coherence of the conversation.
  - Suggest coping techniques based on available knowledge.
  - Provide professional references and support.
- Non-Functional Requirements:
  - Empathy and Support: The language and tone of the chatbot must be empathetic, understanding, and non-judgmental.
  - Security and Privacy: Give users anonymity and total protection of personal information and chat history.
  - Reliability and Accuracy: Information presented must come from a credible source and established scientific fact.
  - Ease of Use: User-friendly, simple, and intuitive interface for all users.
  - Responsiveness: Reduce latency to ensure continued interest.
  - Consistency: Provide consistent user experience between sessions.
- Interaction Principles:

- Create a safe, non-judgmental environment.
- Promote open self-disclosure.
- Give users control of their conversation and personal information.

The output of the user requirements analysis will be a critical input document that will inform the design decisions regarding system architecture, user experience, language model selection, and tuning practices. This is necessary to help guarantee that the chatbot is designed in such a manner as to best address actual needs of the users in the domain of mental health care.

A deep and comprehensive understanding of users' needs, expectations, and psychological barriers is a prerequisite for the success and usefulness of a specialized chatbot supporting mental health. Given that users approach chatbots with sensitive and emotionally diverse concerns and issues, requirements analysis must go beyond purely technical features and delve into psychological, emotional aspects and desired user experiences. The goal of this phase is to clearly and fully define functional requirements (what the chatbot should do), non-functional requirements (what the chatbot should be like), and core interaction principles. These requirements will serve as a guiding principle, guiding the entire process of designing the RAG system architecture, selecting the retrieval method, the goal of LLM fine-tuning, and designing the user interface, aiming to create a chatbot that can build trust, bring comfort, and provide real support value to users.

This study identifies key user requirements primarily through the synthesis and in-depth analysis of scientific works, articles, and related studies that have been detailed in Chapter 2. Key findings from the literature review, including the need for anonymity, empathetic responses, trustworthy information, 24/7 accessibility, and easy-to-use interfaces, will be systematized into specific requirements. Focusing on target user groups such as those with mild to moderate stress or anxiety, or those seeking initial information about mental health, will help to specify the interaction scenarios and types of support that the chatbot should prioritize. The list of functional, nonfunctional requirements and interaction principles detailed in the previous draft will serve as a guide for the next stages of development.

3.3. System Architecture

The chatbot system for mental health is developed on the Retrieval-Augmented Generation (RAG) architecture, which pairs the capacity to retrieve information from a knowledge base with the natural language generation capability of a large language model (LLM). The system guarantees that the chatbot offers empathetic, correct, and updated responses without needing to retrain the entire model.

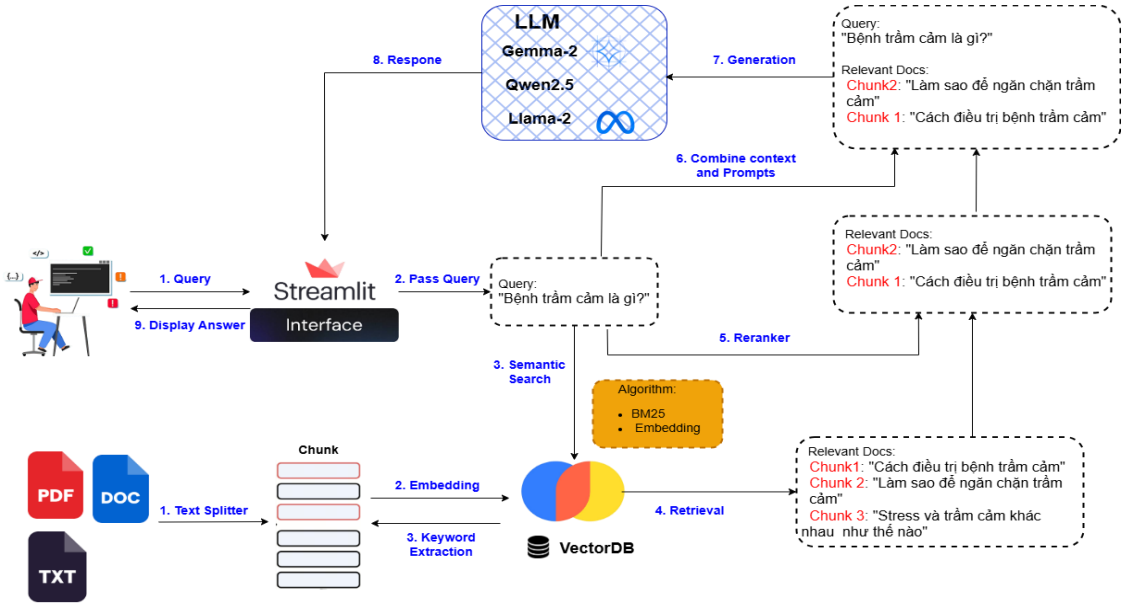

Figure 1. Chatbot system architecture

Figure 1 illustrates the workflow of a mental health chatbot system based on the RAG architecture. Users enter questions through the Streamlit interface, and the query is processed using a semantic search mechanism in a vector database. Relevant text segments are extracted, rearranged, and combined with the query to send to a large language model (LLM) such as Gemma-2, Qwen2.5, or Llama-3 to generate accurate, contextually relevant, and emotionally rich responses. This process helps the chatbot maintain its intelligent response capabilities without retraining the entire model, while ensuring flexibility and updating from the input knowledge base.

❖ RAG-Based Model for Mental Health Support

To deal with the intricacies that come with delivering thoughtful and context-dependent mental health assistance, our architecture incorporates a state-of-the-art Retrieval-Augmented Generation (RAG) approach. This approach takes advantage of the strengths of Large Language Models' (LLMs), while grounding their output in an ever-increasing database of mental health content. To ensure the highest relevance and accuracy of the retrieved information, our RAG system implements a multi-step retrieval process, starting with retrieving potential candidates using both the keyword-based BM25 algorithm and embedding-based semantic search. Next, a dedicated reranker model screens and re-ranks these candidates to select the most optimal context. This allows the chatbot to provide more accurate and relevant responses, effectively mitigating the need for the time-consuming and resource-intensive task of retraining the entire model. The RAG framework is particularly well-suited for tasks that require a lot of knowledge, offering a number of significant advantages such as enhanced accuracy through the use of external knowledge, access to up-to-date information, inherent transparency in the reasoning process, flexibility to specific domains, robust security, high reliability, and easy scalability. The intricate operations of every stage of our RAG model, as depicted in **Figure 1**, will be explained in the subsequent subsections.

❖ Document Preprocessing and Knowledge Indexing

The initial phase of our RAG pipeline is meticulous document preprocessing to transform raw mental health data into a readily structured and easily searchable format. As seen in **Figure 1**,

the initial step begins with a Text Splitter, which divides the input.txt files into small, manageable fragments known as "chunks." This chunking method is required for enhancing retrieval effectiveness as well as for maintaining the relevance and focus of the context presented to the LLM on the user's query. Then these chunks of text are passed through an Embedding operation.

The text of every segment is transformed into a high-dimensional vector representation using the BAAI/bge-m3 model. This model was used due to its high multilingual capability and its ability to capture semantic nuances relevant to mental health contexts. Such vector representations enable fast similarity-based retrieval by preserving the semantic structure of the text.

Concurrently, a Keyword Extraction module can be utilized to extract salient terms from each segment, thereby improving retrieval via an additional lexical match process.

The embedding vectors thus obtained, along with their corresponding textual chunks and keywords, are then inserted into a VectorDB. The system uses Chroma, a vector store system that is highly optimized for similarity search on large embedding collections. Chroma is our primary storage of indexed mental health knowledge and allows us to efficiently update the knowledge repository as new knowledge is created.

❖ Retrieval Phase: Hybrid Candidate Retrieval and Context Reranking

The Retrieval Phase plays a key role in the RAG architecture, responsible for finding and providing the most relevant information from the knowledge base (VectorDB in **Figure 1**) to serve as context for the Large Language Model (LLM) to generate responses. To optimize the quality of context, our system applies a multi-step retrieval process including: hybrid candidate retrieval and context reranking.

1. Hybrid Candidate Retrieval:

When a user enters a query through the interface (e.g. Streamlit as shown in Figure 1), this query is processed simultaneously by two main retrieval methods to generate an initial set of candidate text chunks:

- Vector-Based Semantic Search:
  - The user query is converted into a semantic vector using an embedding model
  - The Semantic Search module then calculates the similarity (e.g. cosine similarity) between this query vector and the embedding vectors of all text chunks stored in VectorDB (Chroma).
  - A top-k list of chunks with the highest semantic similarity is selected.
- Keyword-Based Search (BM25):
  - In parallel with semantic search, the BM25 algorithm (Okapi BM25) is also applied. BM25 is a bag-of-words ranking algorithm that evaluates the relevance of chunks based on the frequency of keywords appearing in the query, adjusted for the length of the chunks and the word specificity in the entire corpus.

- BM25 helps to add chunks that may contain the exact keyword from the query but are not necessarily highly similar in terms of overall semantics.

- A top-k' list of the most relevant chunks according to BM25 is also obtained.

The final set of candidate chunks for the next step is created by combining (i.e., taking a union or merging and removing duplicates) the results from both Semantic Search and BM25. This combined approach aims to leverage the strengths of both methods: the deep semantic understanding of embedding and the keyword precision of BM25.

2. Context Reranking:

After obtaining the list of candidate chunks from the combined retrieval step, a dedicated reranker model is applied.

- A reranker is a model typically based on a Transformer (e.g., Cross-Encoder) that is capable of further evaluating the relevance between each pair (query, candidate chunk).
  • It takes the user's original query and the list of candidate chunks as inputs, and then calculates a new relevance score for each chunk.

- Based on this score, the chunks are reranked more accurately. This step is important to filter out the noisy or less relevant results that the initial retrieval methods may return, and push the truly important chunks to the top of the list.

3. Context Selection for Generation:

Finally, the system selects the top-n from the reranked list. These chunks, with the highest relevance, will be combined with the user's original query to form the final prompt, providing context for the LLM in the Augmentation and Generation phase.

Implementing this multi-step retrieval and fusion process ensures that the LLM receives the richest, most diverse, and most accurate context, thereby improving the quality, consistency, and usefulness of the responses generated for the user.

❖ Augmentation and Generation Phase: Contextualized Response Generation

The most pertinent document pieces, retrieved during the retrieval phase, are subsequently merged with the initial query of the user in the Context and Prompt Combination phase. This merged input constitutes the final prompt that is provided to the large language model (LLM).

The system utilizes state-of-the-art LLMs such as Qwen2.5, Gemma2, or Llama3, which have shown great performance in natural language generation and comprehension.

They have been chosen specifically due to their capacity for understanding intricate contexts and delivering contextually relevant, empathetic, and coherent responses—qualities valuable to applications in mental health care. For example, Llama3 is used with preference due to its high degree of contextual sensitivity and its capacity for generating human-like and empathetic conversation suitable to the sensitive nature of communication entailing mental health.

The complete and accurate response is produced by the large language model (LLM) after taking the synthesized input. By grounding the output in retrieved data, the system makes sure that the responses provided by the chatbot are not only coherent and natural-sounding but also factually accurate and explicitly pertinent to the question regarding mental health asked by the user. The system's output becomes much more credible and helpful with this retrieval-augmented generation (RAG) strategy.

Lastly, the response generated is displayed to the user through the Display Answer interface, completing the chatbot interaction cycle. The modular design allows the chatbot to be flexible and scalable as it does not involve repetitive retraining of the LLM for the integration of new data or evolving user requirements.

3.4. System Evaluation Metrics

This section outlines the evaluation metrics and experimental design approaches used to assess the performance of the AI-based mental health chatbot system, specifically focusing on the Retrieval-Augmented Generation (RAG) architecture and the fine-tuned Large Language Models (LLMs). These metrics and designs are applied in the experiments detailed in Chapter 4 to ensure a systematic and objective evaluation of the system's retrieval and response generation capabilities.

❖ Evaluation Metrics

Retrieval System Metrics (Experiment 1)

To evaluate the performance of the information retrieval component within the RAG architecture, the following metrics are employed:

- Precision: Measures the proportion of retrieved documents (chunks) that are relevant among the top K results. It reflects the accuracy of the retrieval system in returning relevant information.
- Recall: Measures the proportion of relevant documents retrieved out of all relevant documents available. It indicates the system's ability to capture all pertinent information.
- Mean Average Precision (MAP): Computes the average precision across all queries, providing a comprehensive measure of ranking quality for relevant documents.
- Mean Reciprocal Rank (MRR): Calculates the average of the reciprocal rank of the first relevant document retrieved, emphasizing the system's ability to prioritize the most relevant result.

These metrics are used to compare different retrieval configurations, including BM25, Semantic Search, Hybrid (BM25 + Semantic Search), and Hybrid with a reranker, ensuring the selection of the most effective setup for providing context to the LLM.

LLM Fine-tuning Metrics (Experiment 2)

To assess the efficacy of fine-tuning the LLMs (Qwen-2.5, Llama-3, Gemma-2) on a mental health-specific dataset, the following natural language processing metrics are used:

- ROUGE (Recall-Oriented Understudy for Gisting Evaluation):

- ROUGE-1: Measures unigram overlap between generated and reference texts, evaluating basic vocabulary coverage.
- ROUGE-2: Measures bigram overlap, assessing coherence and sentence structure.
- ROUGE-L: Measures the longest common subsequence, capturing sentence order and overall meaning.
- BLEU (Bilingual Evaluation Understudy): Evaluates the similarity of generated text to reference text based on n-gram overlap (unigrams to 4-grams), focusing on lexical and structural accuracy.
- METEOR (Metric for Evaluation of Translation with Explicit ORdering): Considers word overlap, synonyms, stemming, and semantic matching, making it suitable for evaluating fluency and contextual relevance in mental health responses.

These metrics provide a quantitative assessment of the LLMs' ability to generate empathetic, contextually appropriate, and therapeutically relevant responses.

❖ Experimental Design Solutions

Experiment 1: Retrieval System Performance

The first experiment evaluates the performance of different retrieval configurations within the RAG architecture. The design involves:

- Test Dataset: A curated set of queries related to mental health, paired with ground-truth relevant documents from the knowledge base.
- Configurations Tested: Four setups are compared: BM25 (keyword-based), Semantic Search (embedding-based), Hybrid (BM25 + Semantic Search), and Hybrid with a reranker (using BAAI/bge-reranker-v2-m3).
- Evaluation Process: Each configuration retrieves the top K documents for each query, and performance is measured using Precision, Recall, MAP, and MRR. The results guide the selection of the optimal retrieval setup for the final system.

Experiment 2: LLM Fine-tuning Efficacy

The second experiment assesses the impact of fine-tuning LLMs on a mental health dataset. The design includes:

- Dataset: Approximately 15,000 question-answer pairs, split into 90% training (13,500 samples) and 10% validation (1,500 samples), derived from mental health dialogues.
- Fine-tuning Approach: QLoRA is used to fine-tune Qwen-2.5, Llama-3, and Gemma-2, with a single epoch to balance performance improvement and avoid overfitting.
- Evaluation Process: Pre-trained and fine-tuned models are compared using ROUGE, BLEU, and METEOR scores on the validation set to measure improvements in response quality, particularly in semantic relevance and empathy.

These experimental designs ensure a robust evaluation of both the retrieval and generation components, aligning with the research objective of developing an effective mental health chatbot.

CHAPTER 4
IMPLEMENTATION AND RESULTS

4.1. Implementation

This section outlines the detailed implementation of the AI-based chatbot system, covering the system architecture, component integration, and specific configurations used for both the retrieval and language generation modules. It aims to provide a clear understanding of how the theoretical framework translated into a functional prototype.

❖ Development Environment and Tools

- Programming Language: Python, HTML, CSS
- Main Libraries and Frameworks:
  - Natural Language Processing and LLMs: Hugging Face Transformers (for model loading and fine-tuning), peft (for Parameter-Efficient Fine-Tuning with QLoRA), bitsandbytes (for 4-bit quantization), and accelerate (for distributed training).
  - Deep Learning: PyTorch 2.0, used as the backend for model training and inference.
  - Information Retrieval: rank_bm25 (for BM25 implementation), sentence-transformers (for embedding generation with BAAI/bge-m3 and reranking with BAAI/bge-reranker-v2-m3), and scikit-learn (for TF-IDF and other preprocessing utilities).
  - Vector Database: Chroma, for storing and querying embedding vectors.
  - Conversation History Database: MongoDB, for managing user interactions.
  - User Interface: Streamlit, for building a web-based interactive interface.
  - Data Processing: Pandas and NumPy, for handling and cleaning datasets.

❖ Data Collection and Preprocessing

1. Knowledge Base Corpus for RAG:

- Brief description of data sources: The data for the RAG knowledge base is collected from a single file in .txt format, containing selected information on the topic of mental health from various online sources. The information segments in this file are clearly separated by # signs in the structure as shown in **Figure 5**.

Câu 1: Trầm cảm là gì?

Trầm cảm là một rối loạn tâm thần phổ biến. Nó ảnh hưởng đến cảm xúc và suy nghĩ. Các triệu chứng bao gồm buồn bã kéo dài, mất hứng thú, thay đổi giấc ngủ.

Nguồn: https://hellobacsi.com/

Câu 2: Bệnh tâm thần ở Việt Nam

Theo thống kê của Bộ Y tế Việt Nam năm 2023, cả nước có đến 14 triệu người mắc bệnh rối loạn tâm thần hay còn có thể gọi là bệnh tâm lý. Trong đó bao gồm nhiều rối loạn tâm thần khác nhau như trầm cảm, sang chấn tâm lý, rối loạn tâm thần, rối loạn lo âu, rối loạn ám ảnh cưỡng chế, tâm thần phân liệt.

Nguồn: https://moh.gov.vn/

Câu 3: Rối loạn nhân cách chống đối xã hội và tính cách bốc đồng là gì?

Rối loạn nhân cách chống đối xã hội (Antisocial Personality Disorder – ASPD) là một dạng rối loạn sức khỏe tâm thần khiến cho một cá nhân có khuynh thướng thực hiện các hành vi, mà có thể dẫn đến hậu quả nghiêm trọng, nhưng không hề cảm thấy hối hận. Họ có thể tỏ ra thiếu tôn trọng người khác, hung hăng, bạo lực hay thậm chí là liều lĩnh.

Nguồn: https://www.msdmanuals.com/

Figure 5. The format of data.txt for RAG

- Pre-processing details:
  - Text Splitting: Since the original data has been split into separate chunks according to each question and related information by the # sign, the text splitting process uses this delimiter-based splitting method. Each chunk is located after the # sign. With the current data structure, the size of each chunk will correspond to the content of each question and the accompanying source information. The overlap between chunks is not applied in this case because the chunks are clearly separated.
  - Embedding Generation: Each text chunk after being split will be converted into embedding vectors using the BAAI/bge-m3 model. This model was chosen for its multilingual processing capabilities and is particularly effective in capturing mental health-related semantics, ensuring that the embedding vectors accurately represent the context of each piece of information.
- Number of documents, total chunks after processing:
  - Number of original documents: 1 .txt file.
  - Total chunks after processing: 248 chunks.

2. Fine-tuning Dataset for LLMs

- To fine-tune language models for improved empathy and contextual relevance in mental health dialogues, a custom dataset was constructed from multiple reliable sources. This dataset includes anonymized transcripts of counseling and psychotherapy sessions, filtered dialogues from the DailyDialog dataset (focused on mental health topics), and public mental health forum posts with personally identifiable information (PII) removed.experiences and discussions occur. All

information from such forums undergoes strict filtering and anonymization to ensure total confidentiality to the users.

- Preprocessing Steps:
  - Cleaning: Employed regular expressions and manual filtering to sanitize PII, special characters, and non-required content.
  - Formatting: Transformed raw text into question-answer pairs, where each sample was a user query and a corresponding reference response.
  - Quality Control: Hand-checked 10% of the dataset to ensure the quality, therapeutic value, and emotional suitability of the responses.
- Dataset Statistics:
  - Total Size: approximate 15,000 QA pairs
  - Split: 90% for training (13,500 samples) and 10% for validation (1,500 samples)

This dataset serves as the foundation for fine-tuning large language models, aiming to enhance their ability to engage in supportive and contextually accurate mental health-related conversations.

❖ RAG System Implementation

The Retrieval-Augmented Generation (RAG) architecture integrates information retrieval and response generation to produce contextually grounded, accurate, and empathetic answers. The system consists of three core components:

1. Knowledge Base Indexing

- Text from the knowledge base is divided into manageable chunks and embedded using the BAAI/bge-m3 model. These embeddings, along with associated metadata such as chunk ID, source, and timestamp, are stored in the Chroma vector database.
- Chroma is chosen for its efficient similarity search and scalability, enabling real-time updates without the need to re-index the entire corpus.

2. Hybrid Retriever

This component combines traditional keyword matching and semantic search to improve retrieval quality:

- BM25 is implemented using the rank_bm25 library, with parameter settings: k_1 = 1.5 for term frequency scaling and b = 0.75 for document length normalization. It provides robust keyword-based alignment between query and text.
- Semantic Search utilizes BAAI/bge-m3 embeddings to calculate cosine similarity between the query and each document chunk, capturing deeper contextual relationships.

- Hybrid Retrieval Strategy retrieves the top 10 results from both BM25 and Semantic Search. Results are merged and deduplicated based on chunk IDs.

- Reranking is performed using BAAI/bge-reranker-v2-m3, which scores the merged results and selects the top 5 most relevant chunks to be used as context.

The final output of this component is a set of five top-ranked text chunks, concatenated with the user query to form the full prompt for the generation module. The retrieval pipeline is built with custom Python code using rank_bm25, sentence-transformers, and Chroma.

3. Generation Module

This component is responsible for generating empathetic and relevant responses:

- In Experiment 1, the baseline response generation is handled by the Qwen-2.5 3B-Instruct model (4-bit quantized).

- In Experiment 2 and the final system, a fine-tuned Llama-3 model is used, selected due to its superior performance in METEOR evaluation.

- Prompt Construction involves combining the user query with retrieved context, structured with specific instructions to promote empathy and factuality in the generated response.

❖ LLM Fine-tuning Implementation (QLoRA)

To improve the performance of large language models (LLMs) in mental health-related conversations, three models, Qwen-2.5, Llama-3, and Gemma-2, were fine-tuned using the Quantized Low-Rank Adaptation (QLoRA) method. This method helps the model learn more efficiently by updating only a small portion of the parameters, thereby saving memory and training time.

Refined models:

- Qwen-2.5: 3 billion parameters, 4-bit quantization

- Llama-3: 3 billion parameters, 4-bit quantization

- Gemma-2: 2 billion parameters, 4-bit quantization

Fine-tuning Method: QLoRA, which adapts pre-trained models efficiently by updating a small subset of parameters (low-rank adapters).

QLoRA Configuration:

- Rank (r): 8

- LoRA alpha: 16

- LoRA dropout: 0.1

- Target modules: ["q_proj", "k_proj", "v_proj", "o_proj", "gate_proj", "up_proj", "down_proj"]

Hyperparameters:

- Learning Rate: 2e-4.
- Batch Size: 4 (gradient accumulation steps = 4 for effective batch size of 16).
- Epochs: 1, to balance performance improvement and prevent overfitting.
- Optimizer: AdamW with weight decay = 0.01.

Training Data: 13,500 question-answer pairs from the fine-tuning dataset.

Validation: 1,500 samples used to monitor performance and prevent overfitting.

4.2. System Design Details

This section elaborates on the key design components of the AI-based mental health chatbot system, including the database structure, user interface, and ethical considerations. These elements ensure the system's functionality, usability, and alignment with ethical standards.

❖ Database Design

The system employs two types of databases to support efficient information retrieval and user interaction management:

1. **Vector Database (Chroma)**

   - Chroma acts as our state-of-the-art vector storehouse, responsible for the storage of embedding vectors of processed pieces of mental health information. The database is optimized with the sole aim of providing swift and accurate similarity searches to support the system to efficiently retrieve the most relevant contextual information in response to user questions.

   - Chroma's internal architecture is basically comprised of embedded vectors, corresponding text chunks, and relevant metadata (i.e., the source). Another advantage of utilizing Chroma is that it has a dynamic update mechanism, meaning that updated or new information can be added without any hiccup. As mental health information changes over time, Chroma allows for dynamic updates to the stored vectors, thereby enabling the chatbot to always draw on an updated and accurate knowledge base.

   - This capability is paramount within the world of mental health, where accuracy, punctuality, and moral responsibility are intrinsic to maintaining user trust and being able to deliver appropriate support.

2. **Conversation History Database (MongoDB)**

   - To facilitate personalized interactions and enable the tracking of user progress, MongoDB is employed as the primary database for storing user-specific data. As a NoSQL document-oriented database, MongoDB is particularly well-suited for handling the flexible and evolving nature of conversational data inherent in chatbot systems.

- Specifically, MongoDB is used to persist the conversation history between each user and the chatbot. Each interaction—including the user's queries and the chatbot's responses—is stored as a document associated with a unique user identifier. This design allows the system to:

    - Maintain context across multiple turns: The stored conversation history can be retrieved and included as context for subsequent interactions, enabling the chatbot to provide more coherent and relevant responses.

    - Personalize the user experience: By referencing past interactions, the chatbot can tailor its responses to better align with the individual user's concerns, preferences, and communication style.

    - Monitor user progress and identify trends: Over time, aggregated and anonymized conversation data can be analyzed to detect behavioral patterns or recurring issues, offering insights for improving the system's effectiveness and mental health support strategies.

- To ensure scalability and adaptability, the MongoDB schema for conversation logs is designed to be flexible, commonly including fields such as:

    - user_id: A unique identifier for each user.

    - timestamp: The time of the interaction.

    - user_message: The user's input text.

    - bot_response: The chatbot's reply.

    - session_id (optional): Identifier for a specific conversation session

- This approach supports both the technical demands of dynamic conversational systems and the ethical responsibilities required in delivering mental health support through AI.

❖ User Interface Design

This section presents the design considerations and decisions for the user interface (UI) of a mental health support chatbot. The goal is to create an interface that is intuitive and easy to use while ensuring confidentiality, trustworthiness, and sensitivity to the mental health field. The UI design is user-centered, focusing on providing an effective and comfortable support experience.

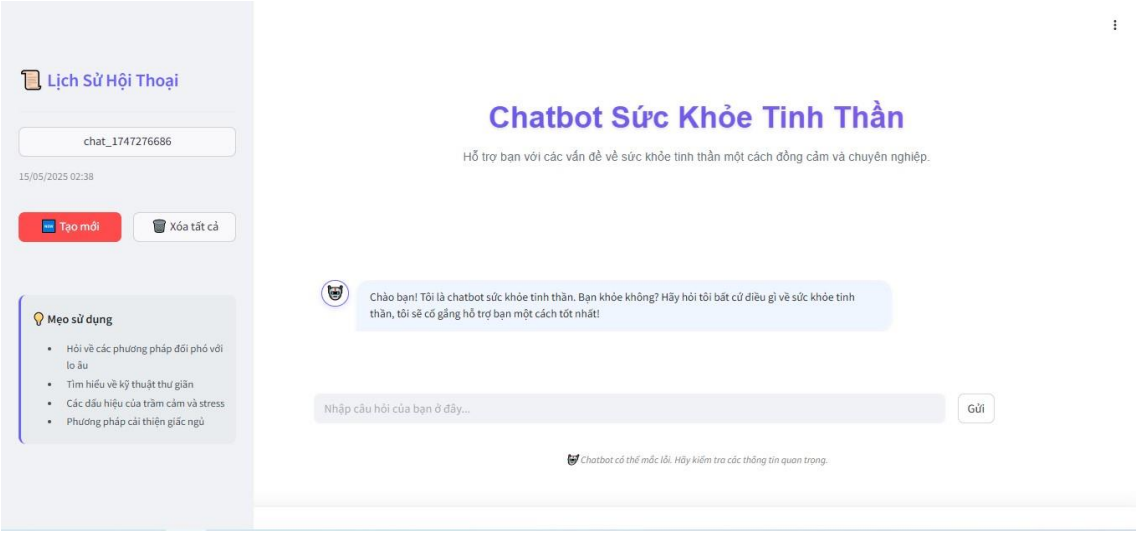

Figure 2. Illustrate the main interface of the chatbot

Design Principles: Figure 2 is the chatbot's user interface is built on the following design principles:

- Simplicity and Clarity: A clean layout, simple terms, and lack of complex technical phrases characterize the interface design. Interactive components appear in viewable form in order to enable users to understand in a straightforward manner how to utilize the chatbot without ambiguity. For example, the chatbot would foreground presentation of a prominent chat window with visible input fields and minimal function buttons.

- Intuitiveness: Interaction flow is reasonable and predictive. Users are able to navigate intuitively and get the info or functionality needed without being overly relentlessly directed. Symbols and icons (if present) are always used and express well-known meaning. Home page of chatbot clearly indicates a welcome and usage guide so that a fresh user is easily able to start.

- Consistency: Consistency is maintained across the entire interface in terms of visual style (e.g., color, typography), language used, placement of interactive elements, and system responses. This helps users build familiarity and ease of interaction with the chatbot throughout their use. Elements such as the page title “Mental Health Chatbot” and input fields remain consistent across different pages.

- Friendliness and Supportiveness: The chatbot’s communication language is carefully chosen to be encouraging, empathetic, and non-judgmental. Messages and responses are designed to provide a sense of support and reassurance to users, especially when they are sharing about sensitive issues. The chatbot greets users in a friendly manner and encourages them to share.

- Privacy and Insecurity: The user interface design indirectly safeguards user data by preventing unnecessary exposure of private data in normal usage, though a "Conversation History" function is set up to allow users to view their previous conversations so that they can utilize this function, implying controlled data storage.

- Accessibility: While it may not be a primary focus at the early stages, UI design still considers basic accessibility factors, such as ensuring sufficient color contrast for

readability and using appropriate font sizes. More comprehensive accessibility improvements may be considered in later stages of development.

❖ Key Interface Components

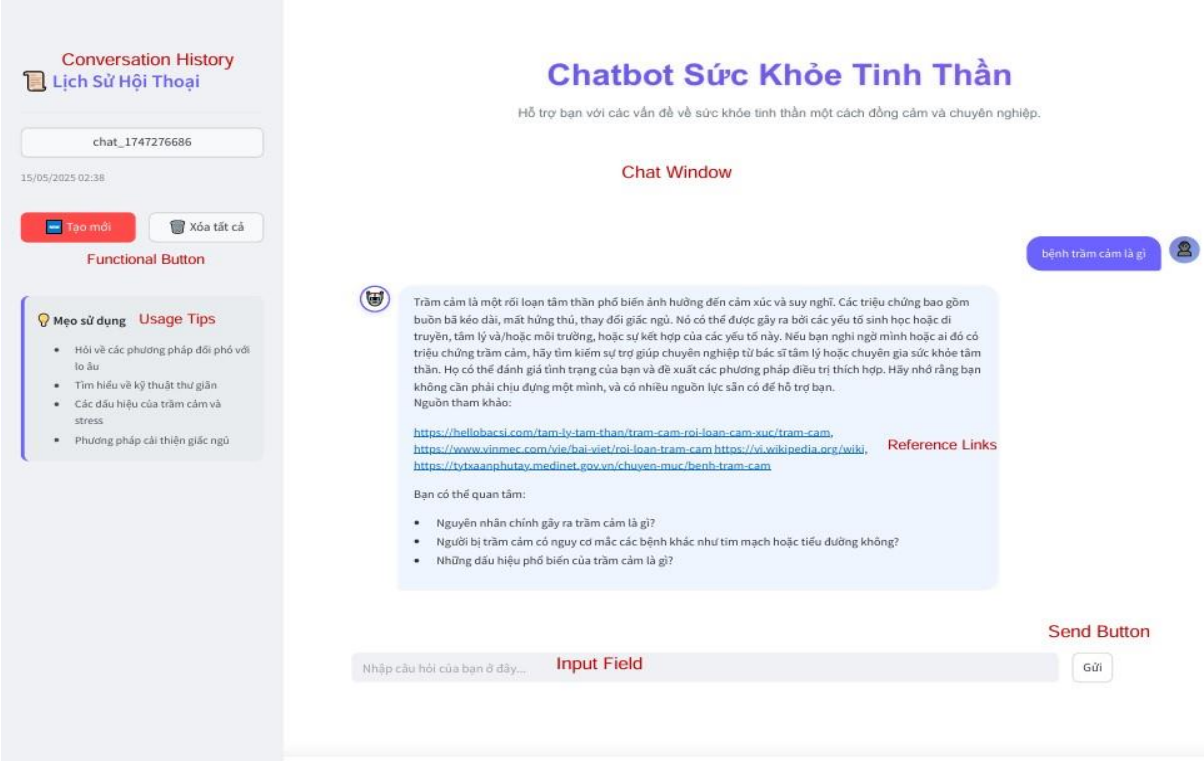

Figure 3. Interface layout diagram

Figure 3 shows that the chatbot's user interface consists of the following main components:

- Chat Window: This is the central area that displays the message history between the user and the chatbot. Messages from the user and the chatbot are clearly distinguished. This area is designed to be scrollable, allowing the user to review previous messages.
- Input Field: A single-line text field placed at the bottom of the chat window, allowing the user to enter their question or message. This field has placeholder text such as "Enter your question here..." to guide the user.
- Send Button: A "Send" button placed near the input field, allowing the user to send their message to the chatbot after typing.
- Conversation History: A function that allows the user to review their previous conversations with the chatbot.
- Functional Buttons: The “Create New” and “Clear All” functional buttons allow users to start a new conversation or clear the current conversation history.
- Usage Tips: A “Usage Tips” section provides tips on how to use the chatbot and mental health topics that the chatbot can help with.

- Reference Links: Reference links are provided so users can learn more about mental health topics.

❖ User Interaction Flow

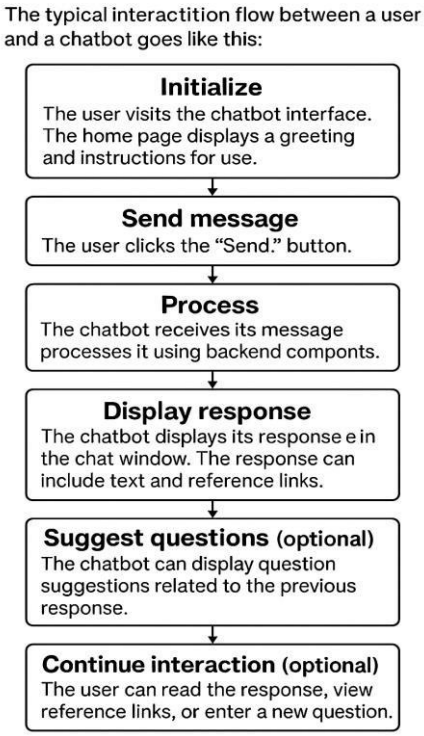

Figure 4. User interaction flow diagram

This is how a user and a chatbot typically interact show **Figure 4**:

- Start: The user accesses the chatbot's user interface. A greeting and usage instructions are shown on the home page.
- Input: The user fills in the input field with their query or message.
- Send message: The "Send" button is clicked by the user.
- Procedure: After receiving the message, the chatbot uses backend components to process it.
- Display response: The chat window shows the chatbot's response. Both text and reference links may be included in the response.
- Suggest questions: In relation to the prior response, the chatbot may present suggested questions.
- Optional: The user can choose one of the suggested questions, read the response, view reference links, or ask a new one.

This flow is made fluid and intuitive by optimizing the interface design. The submit buttons and input fields are positioned in convenient areas. Users can easily follow the context by viewing the chat history. Reference links and the "Tips" feature facilitate effective information discovery for users.

❖ Design Tools and Technologies

An open-source Python framework called Streamlit, that makes it easy and quick to create interactive web apps from Python scripts, was the primary tool used to create the user interface for this chatbot. Streamlit was chosen because of its ability to:

- Faster user interface development: Streamlit uses a lot less HTML and CSS code than traditional web development, which allows for more time to focus on the core functionality of the application. To achieve the necessary level of interface customization, a very small portion of the HTML and CSS code is also utilized to modify layout, typography, and color schemes.

- Rapid prototyping: Streamlit makes it possible to quickly create and test user interface prototypes, which is particularly helpful in the design and iteration stages.

- Simple Python integration: Streamlit offers smooth Python integration with other projectrelated Python libraries and modules because the chatbot is written in Python.

Streamlit functions and widgets are used to create basic interface elements like chat windows, input fields, submit buttons, and other display elements. In the meantime, the interface is further customized by embedding or combining HTML and CSS code, guaranteeing that it satisfies aesthetic and user experience standards.

4.3. Results

This chapter presents quantitative results collected from two main experiments designed to evaluate and optimize the core components of a mental health care chatbot system based on the Retrieval-Augmented Generation (RAG) architecture. Experiment 1 focuses on the performance of the information retrieval system, and Experiment 2 evaluates the performance of fine-tuning large language models (LLMs) for this specific application domain.

❖ Experiment 1: Retrieval System Performance

The goal of Experiment 1 is to systematically evaluate the performance of different information retrieval conditions within the RAG setup, so that which context is given to LLM is most suitable and precise. Incorporated among these conditions are traditional, semantics-driven, and mixed retrieval modes, with or without the support of a reranker model.

- **Measure of Evaluation:**

  - Precision: Number of relevant documents (chunks) out of the top K results retrieved. This parameter indicates the accuracy of returned results.

  - Recall: Number of relevant documents out of top K results divided by the number of actual relevant documents. This parameter indicates the ability of the system to retrieve relevant information.

- Mean Average Precision (MAP): Average of Average Precision (AP) over all queries, providing a general measure of the ranking quality of relevant documents.

- Mean Reciprocal Rank (MRR): The mean of the reciprocal rank of the first relevant document retrieved. MRR is especially valuable when customers wish to retrieve an accurate answer as soon as possible.

- **Comparative Performance of Retrieval Configurations**

Method	Precision	Recall	MRR	MAP
TF-IDF	0.3448	0.3448	0.6839	0.3046
TF-IDF + Reranker	0.5632	0.5632	0.8908	0.5529
BM25	0.3448	0.3448	0.6092	0.2720
BM25 + Reranker	0.4712	0.4712	0.8046	0.4406
Embedding	0.5747	0.5747	0.9252	0.5115
Embedding + Reranker	0.6436	0.6436	0.9023	0.6034
Hybrid (BM25 + Embedding)	0.6436	0.6436	0.9023	0.6034
Hybrid +Reranker	0.7011	0.7011	0.9655	0.6704

Table 2. Performance Results Retrieved

Table 2 contrasts the performance of various retrieval methods on four standard metrics: Precision, Recall, Mean Reciprocal Rank (MRR), and Mean Average Precision (MAP). The table clearly indicates the impact of reranking and hybrid methods on retrieval quality.

Key observations from the table include:

- Embedding-based methods outperform traditional keyword-based methods (TF-IDF, BM25) on all metrics.

- Reranking significantly boosts performance. Incorporating a reranker into TF-IDF, for example, brings MAP forward from 0.3046 to 0.5529.

- Overall best performing is the Hybrid approach (BM25 + Embedding) with reranker with Precision = 0.7011, MRR = 0.9655, and MAP = 0.6704.

- The boost in both Precision and Recall consistently by rerankers shows how they are able to bring forth more relevant information.

These findings are also graphed in **Figures 6** and **7**, which plot the gains over different methods and further support the effectiveness of reranking in combination with method combination.

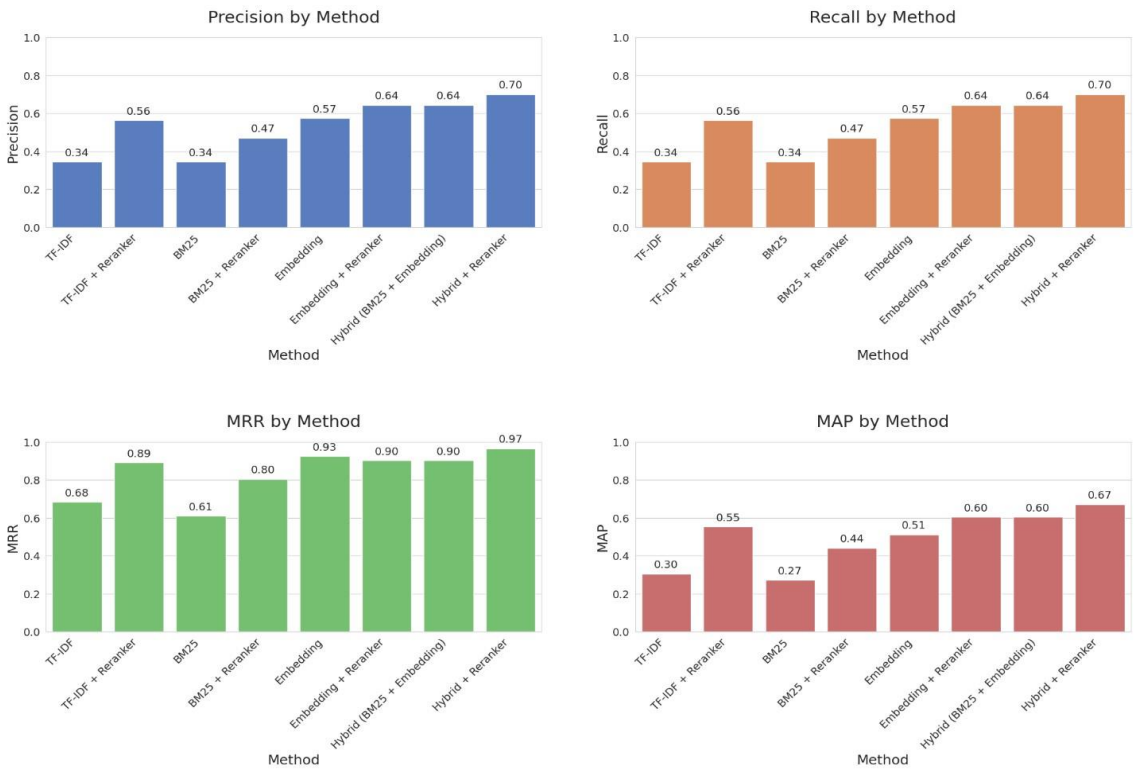

Figure 6. Comparison of Precision, Recall, MAP, MRR of retrieval methods

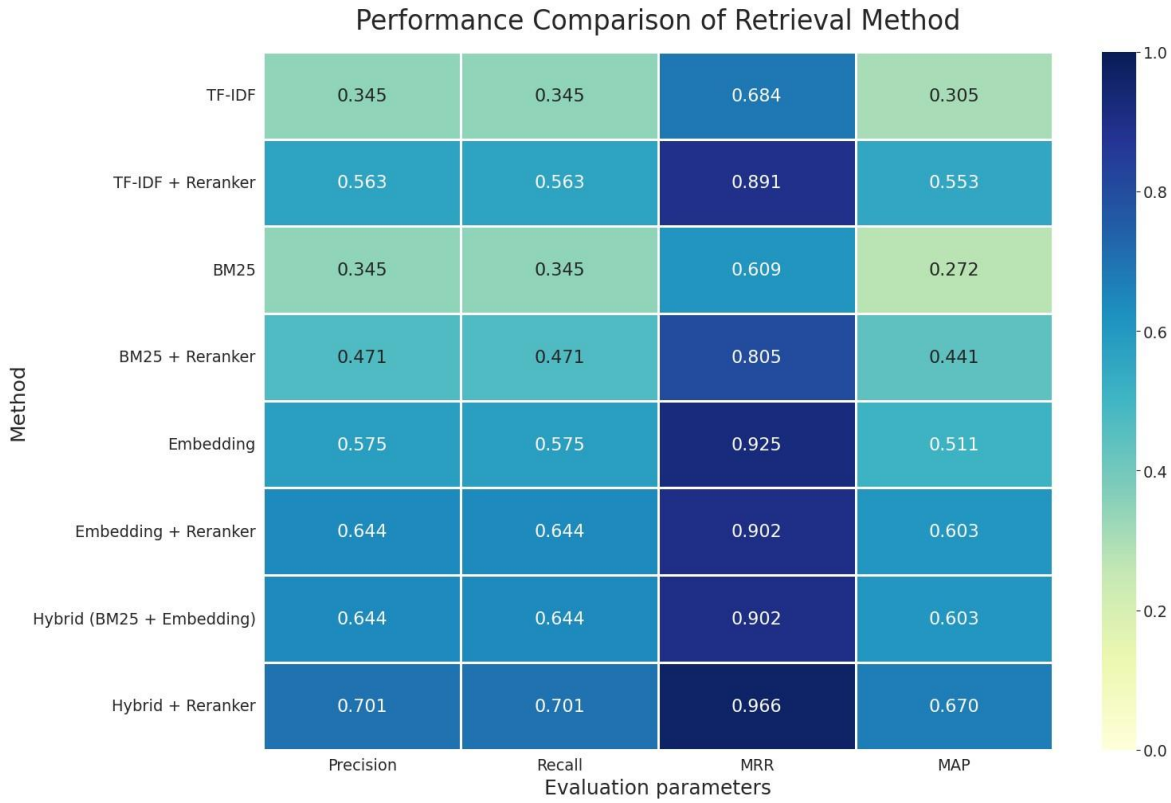

Figure 7. Heatmaps correlate performance between access configurations.

- Optimal Retrieval Configuration Identification:

  - From the data in **Table 2** and the charts (**Figure 6**, **Figure 7**), it is clear that the Hybrid (BM25 + Embedding) + Reranker configuration has demonstrated superior performance on all evaluation metrics. With Precision and Recall reaching 0.7011, MRR reaching 0.9655 and MAP reaching 0.6704, this configuration not only ensures the retrieval of a lot of relevant information but also prioritizes the most important information effectively.

  - This excellence is due to complementarity between the elements: BM25 helps in extracting the exact keywords, whereas the Embedding model provides rich semantic understanding between the query and text segments. The fusion of the two models helps in benefiting from keyword search and semantic encoding. Finally, the Reranking stage is charged with refining and re-ranking potential candidates, based on a more comprehensive evaluation criterion, thereby significantly improving the accuracy and relevance of final results provided to the LLM. Hybrid + Reranker was therefore chosen to be included within the ultimate RAG architecture of the system, for maximizing the context retrieval efficiency as well as improving the response quality of the language model.

❖ Experiment 2: LLM Fine-tuning Efficacy

Experiment 2 was set up to analyze the effect and effectiveness of fine-tuning big language models (Qwen-2.5, Gemma-2, and Llama-3) on a specialized mental health dataset. The aim was to measure the capacity of the models to produce empathetic, contextually adequate, and therapeutically useful responses before fine-tuning and after fine-tuning.

- **Measure of Evaluation:**

  The execution of the fine-tuning procedure is evaluated using popular automatic natural language processing metrics like ROUGE (ROUGE-1, ROUGE-2, ROUGE-L), BLEU, and METEOR. These metrics calculate similarity between the model-generated text and reference text in terms of lexical overlap, word order, and response fluency.

  ROUGE (Recall-Oriented Understudy for Gisting Evaluation):

  - ROUGE-1: Measures the degree of overlap in unigrams (single words) between the reference and generated text, measuring the coverage of the most primitive vocabulary.

  - ROUCE-2: Is directed towards bigrams (adjacent word pairs), measuring coherence and sentence structure.

  - ROUCE-L: Is directed towards the Longest Common Subsequence, measuring the ability to maintain sentence order and overall sentence meaning.

  BLEU (Bilingual Evaluation Understudy): Compares similarity of generated and reference text on the basis of n-grams (from unigrams to 4-grams), with more weight on longer sequences. BLEU is particularly good for analyzing lexical and sentence structure accuracy.

METEOR (Metric for Evaluation of Translation with Explicit ORdering): Considers not only word repetition, but also synonyms, word stems, and semantic matching. METEOR is suitable for judging fluency and contextual sense, especially in medical applications like responding to mental health questions.

- **Performance of Pre-trained LLMs:**

LLM models	ROUGE-1	ROUGE-2	ROUGE-L	BLEU	METEOR
Qwen-2.5	0.5247	0.3044	0.2768	0.0261	0.2267
Gemma-2	0.5849	0.3380	0.3025	0.0594	0.2391
Llama-3	0.5220	0.2931	0.2968	0.0511	0.2371

Table 3. Performance of Pre-trained LLMs

Table 3 illustrates the performance of three language models, namely Qwen-2.5, Gemma-2, and Llama-3, prior to fine-tuning on the test dataset. The used metrics are ROUGE-1, ROUGE-2, ROUGE-L, BLEU, and METEOR, collectively measuring the models' ability to generate relevant and fluent responses compared to reference responses.

Among the three, the pre-trained Gemma-2 model initially performed best on most metrics, achieving the top rank in ROUGE-1 (0.5849), ROUGE-2 (0.3380), ROUGE-L (0.3025), and METEOR (0.2391). Llama-3 surpasses Qwen-2.5 very narrowly in ROUGE-L and METEOR, and Qwen-2.5 lags behind in all metrics—most critically BLEU, where it manages only 0.0261, reflecting that its generated outputs were less aligned with ground-truth references.

These baseline scores constitute a strong rationale for tuning, especially for improving the coherence, fluency, and empathetic character of generated responses—vital features in the context of mental health care.

- **Performance of Fine-tuned LLMs:**

Fine-tuned LLM Model	ROUGE-1	ROUGE-2	ROUGE-L	BLEU	METEOR
Qwen-2.5 (Fine-tuned)	0.5305	0.3192	0.2856	0.034	0.2312
Gemma-2 (Fine-tuned)	0.5816	0.3350	0.3198	0.0612	0.2382
Llama-3 (Fine-tuned)	0.5318	0.3089	0.3060	0.0541	0.2453

Table 4. Performance of Fine-tuned LLMs

Table 4 shows the performance of the models after fine-tuning on a mental health dataset of approximately 15,000 dialogue examples. All three models show improvements in most of the metrics from their original performance. Specifically, Llama-3 (Fine-tuned) achieves the highest METEOR score (0.2453), which indicates the most improved semantic alignment and fluency in its responses. Gemma-2 continues to top the ROUGE-L, which reflects its ability for

preserving structure and relevance. These results confirm that fine-tuning enhances the models' capacity to generate more accurate and empathetic responses in mental health conversations.

Some of the gains on the metrics after the fine-tuning procedure are noted from Tables 2 and 3. For instance, Llama-3 (Fine-tuned) achieved a METEOR score of 0.2453, an improvement over the pre-trained counterpart (0.2371) and best among the fine-tuned models. This shows that finetuning has helped the model generate responses that are semantically closer and more pertinent to the specialized discourse of the mental health field. Qwen-2.5 also showed improvement in most of the ROUGE and BLEU scores upon fine-tuning. Gemma-2, although showing a slight decrease in ROUGE-1 and ROUGE-2 scores, showed improvement in ROUGE-L and BLEU. These changes reflect that fine-tuning, even with one epoch, has begun to specialize the models to mental health conversation, but the degree of improvement and stability across measures is uneven for the models.

The results of Experiment 2, which were done to optimize large language models (LLMs), showed non-uniformity in performance. While Llama-3 impressed with substantial improvements on the METEOR metric and subjective feedback quality, the other models (Qwen-2.5, Gemma-2) did not show even improvement on all metrics, and even showed marginal degradation in some points.

There can be many explanations for this. Second, the size and nature of the fine-tuning data may not be large or representative enough to allow the other models to finish converging and optimizing on all the metrics. Different baseline models will also contain different features that would cause them to respond differently to identical fine-tuning data. Additionally, using only one fine-tuning epoch may prove to be too little to achieve complete convergence in all dimensions, especially for larger sets of data or more complex tasks. Finally, there are incompatibilities between the finetuning data and the dimensions that some metrics are being measured across such that gains in some of said metrics won't equate to improvement in others, or even minor regression as the model learns to a set of objectives. Llama-3 may have had a more ideal shape or fit to our specific finetuning data and objectives that resulted in improved performance.

Summary of Main Results: The results of these two experiments have significant implications for the development of AI chatbots for mental health support:

- Hybrid retrieval system (Hybrid + Reranker) performed better to produce appropriate and relevant context to LLMs, which is a prerequisite for chatbots to output quality responses.

- Fine-tuning of LLM with mental health information was discovered to possess the capability of improving response quality, specifically the capacity to generate more semantically and contextually relevant responses (measured with the METEOR score). Optimal model out of models considered was fine-tuned Llama-3 on a variety of important metrics. The findings emphasize the importance of both RAG architecture tuning and large language model tailoring in developing good and reliable mental health care systems.

CHAPTER 5
DISCUSSION

This chapter is also devoted to the additional explanation of the research results presented in Chapter 4, comparison to analogous scientific papers, evaluation of the contributions, advantages, and disadvantages of the developed AI chatbot system for psychology and proposing future research directions.

5.1. Summary and Interpretation of Key Findings

The present study successfully designed and evaluated an AI chatbot system with RetrievalAugmented Generation (RAG) architecture and fine-tuned large language models (LLMs) to provide empathetic and contextually appropriate initial mental health treatment.

- Regarding the performance of the retrieval system (Experiment 1): The results unequivocally confirmed the superiority of the Hybrid (BM25 + Embedding) + Reranker configuration. This environment had the maximum Precision (0.7011), Recall (0.7011), MRR (0.9655) and MAP (0.6704) scores. This signifies delivering accurate and varied contexts to LLM, one of the most crucial aspects in generating high-quality, reliable answers, achieving the mission of providing accurate information in a delicate field such as mental health. Efficient utilization of keyword search, semantic search and reranker's fine-tuning ability partially mitigated the problem of preserving information timeliness and accuracy without frequently retraining the entire LLM.

- The efficacy of fine-tuning LLM (Experiment 2): Fine-tuning of Qwen-2.5, Gemma-2, and Llama-3 models on a mental health-specific dataset (circa 15,000 samples) showed some enhancement in the ability to generate contextually relevant text. Most notably, the METEOR score of 0.2453 was realized by the fine-tuned Llama-3 model, showing promise to generate more fluent and semantically rich content in mental health discussions. Although the ROUGE and BLEU scores also varied, the METEOR score is vital because it calculates semantic and phrasing similarity that is essential for such discussions that entail subtlety as well as empathy. These results are in direction toward the comparison and contrast of LLM performance in mental health chatbot systems.

Overall, these results have accomplished the fundamental research objectives: developing a RAG chatbot system from fine-tuned LLMs, evaluating the performance of the LLMs, and measuring the efficiency with the classical metrics.

5.2. Detailed Discussion of Research Results

❖ Efficiency of the Combined Retrieval System in RAG Architecture

The success of the Hybrid (BM25 + Embedding) + Reranker is not coincidental and can be attributed to several key factors. BM25, through its ability to recall some terms and keywords, ensures that information containing vital keywords in the search query ranks higher. At the same time, the use of the embedding model (BAAI/bgem3) allows the system to understand the deep meaning and intent of the query even when the user is not using the precise keywords in the knowledge base. This compilation sets up a preliminary list of candidates that is wide but extremely pertinent.

The reranker model's (BAAI/bge-reranker-v2-m3) function is then completely vital. It not only combines the output, but also re-evaluates the relevance of each candidate text snippet to the original query in a more nuanced way, having the effect of minimizing noise and allowing the truly useful information to take center stage. It is especially vital when it comes to mental health, where providing misleading or unhelpful information can be hazardous.

This result is in accordance with most research that has evidenced the strengths of RAG in improving the response quality of LLMs by supplying stable and background knowledge **[35] [40]**. This research contributes additional empirical evidence on the ability of a specific RAG model, optimized via multiple steps, for use in mental health chatbots.

❖ Impact and Significance of Fine-tuning LLM

The fine-tuning of the large language models (LLMs) in Experiment 2 revealed striking differences between the models tested. Importantly, fine-tuning with a single training epoch produced dramatic improvements for the Llama-3 model, particularly on the METEOR measure and overall quality of perceived response. For the Qwen-2.5 and Gemma-2 models, however, the improvement was non-uniform across all metrics, including minor performance degradation in some areas.

Training has been confined to one epoch, thereby intentionally providing baseline performance, thereby risking minimal or no overfitting. Having a fine-tuning dataset of around 13,500 examples, training for many epochs over a fairly small fine-tuning dataset for an extensive LLM will lead to memorization to training data, thereby generalizing poorly to new data. Whereas a single epoch is sufficient for models to begin to learn particular patterns and subtleties in mental health data, and that was shown to be effective in Llama-3 training with additional epochs being able to possibly lead to improvements and enhancements that are better and more uniform across metrics, where the model gets deeper into converging on the nature of the dataset. This is, however, accompanied by a greater risk of overfitting, particularly as the model will start to memorize certain training examples instead of learning general characteristics. Chasing an initial epoch is hence regarded as a practical compromise between attaining improvements in performance and preventing rote memory.

Optimization and analysis of the training epoch and exploring regularization techniques such as dropout or early stopping is one major potential research area in the future. This will further enhance the fine-tuned model performance and offer improved generalization to actual real-world data.

❖ Compare with Related Projects

The system so developed in this work has a number of objectives in common with previous efforts at using AI and LLM to augment mental health. A number of studies have shown the capability of chatbots in enhancing access to care, reducing stigma, and providing instant assistance **[38] [39]**.

- RAG architecture in mental health support: Use of RAG in this thesis is similar to solutions such as SouLLMate, which also makes use of RAG and prompting techniques for improving the support quality. This thesis explores more deeply into evaluating and optimizing aspects of the retrieval module itself (hybrid retrieval, reranking) and comparing the performance of newer open-source LLMs after fine-tuning, with more detailed technical information. Studies such as OnRL-RAG **[35]** examine RAG

personalization with reinforcement learning, one potential direction not covered by existing systems.

- LLM evaluation and fine-tuning: Even as studies like Mental-LLM focus on mental health condition prediction from text-based internet data, this thesis focuses on interactive response generation capacity in chatbots. Systematic reviews **[38] [39]** typically state that although AI chatbots are promising, making them safe, empathetic, and effective is a challenge. This thesis attempts to surmount some of that challenge by aiming for information trustworthiness and language appropriateness. But the metric of empathy is still largely based on machine-driven language measures, with no real-world test by endusers, an aspect that many studies are also equally seeking to rectify.

❖ Significance and Contribution of the Study

This study has several important implications and contributions:

- Technical: Document empirical findings on how multi-tier RAG (hybrid retrieval + reranker) architecture enhances the quality of information retrieval across expert domains. At the same time, provide an early performance comparison of the impact of nextgeneration open source LLMs (Qwen-2.5, Llama-3, Gemma-2) after fine-tuning on the mental health domain.

- Implementation: Propose a feasible approach to the creation of early mental support chatbots that can deliver reliable information and offer better responses. This might help address the dearth of professionals and cost and geographical barriers in getting access to mental health care services.

- For researchers: Sharing experience and results in data processing, LLM tuning, and system evaluation which can be used as future research references at the intersection of AI and mental health.

❖ Limitations of the Study

Despite the favorable results, there are certain limitations of this research that need to be valued:

- Diversity and size of data: Data employed for RAG (248 chunks) and fine-tuned LLM (approximately 15,000 samples) were carefully chosen and preprocessed but were quite small in size and inadequate in covering a broad spectrum of mental illness. Richness and quality of the data will influence the performance of both RAG and fine-tuned LLM.

- Empathy and measurement of user experience: The study relied primarily on machine measurements of language quality. Fundamental aspects such as the true empathy of the chatbot, user acceptance, and actual therapeutic impact in the real world were not assessed through actual user or psychologist tests. This is one of the weaknesses of most of the existing AI chatbot research, since it has been demonstrated that standard NLP measures are not adequate to assess the therapeutic components of communication.

- Limited scope of LLM: This study only considered three specific LLM models. There are many other commercial as well as open source models available, which can provide different performance levels.

- Computational resources: Such large LLM fine-tuning and execution require enormous computational resources that may be untenable for deployment at scale.

- Ethical issues: While the study assumed the use of ethical controls, the actual release of a mental health chatbot must take care to fully mitigate issues of privacy, data security, potential for spreading misinformation, and liability when the chatbot cannot handle crisis situations.

CHAPTER 6
CONCLUSION AND FUTURE WORK

6.1. Conclusion

This research successfully developed and evaluated a specialized chatbot system capable of addressing mental health questions. This is not merely an intellectual exercise but a pragmatic step towards the application of technology in the field of public health care.

Our major achievements are:

- Reaching a full-scale chatbot system: A robust chatbot framework has been successfully deployed, capable of handling and processing user questions about mental health. The system is technology-driven, ensuring smooth and stable performance.

- Achieving the interactivity and information objectives: The chatbot has reached the stage of understanding and responding to user questions appropriately and in a useful manner. The chatbot not only provides information but also offers relevant suggestions and guidelines, thereby contributing to mental health awareness.

- Contribution to information access for mental health: This project makes an important contribution towards offering a simpler and more anonymous platform to access mental health information. Psychological support remains a barrier for most people, and this chatbot may serve as an initial support that is capable of de-stressing and offering more self-learning potential.

- Demonstrating the practical potential of AI in medicine: The current study convincingly demonstrates the huge potential of artificial intelligence in medicine, and especially in mental healthcare. It presents new ways for creating intelligent assistance tools and assisting in overcoming existing issues in the health system.

- Developing the foundation for future research: The results and learning from this project will build a strong foundation for future R&D. Opportunities for incorporating additional depth to the chatbot's knowledge base, incorporating more sophisticated support features, or researching more complex interaction modes to enhance the user experience are possible.

In conclusion, through In conclusion, this project not only fulfilled its academic requirements but also contributed to the application of technology in addressing a significant social problem. It is believed that this product will have practical applications and will provide a foundation for future innovations.

6.2. Future Work

Apart from further expansion and improvement of the efficiency of this chatbot system, a number of future research and development directions are possible. First of all, future work will focus on expanding and improving the knowledge base of this chatbot. This means adding more authoritative medical information sources, continually updating the latest knowledge with regard to mental health, and improving subjects to cover more aspects of the issue.

Second, a key area for future improvement is the natural language understanding (NLU) and natural language generation (NLG) ability of the chatbot. Further research should explore the

use of more advanced deep learning models, specifically large language models (LLMs), to improve the accuracy level of interpreting user intent and generating more natural and responsive answers. This will make the chatbot engage more personally and efficiently with users, and minimize cases of misinterpretation or off-topic responses.

Furthermore, incorporating elements of personalized assistance is a viable possibility. Future work could explore personalizing assistance by maintaining user interaction histories to tailor advice more effectively. Also, a partnership with medical professionals or organizations that offer mental health assistance can be visualized in creating a referral system or open consultation if necessary, expanding the scope of assistance being offered by the chatbot.

Finally, there should be more in-depth effectiveness and user satisfaction evaluation studies on the chatbot. These involve soliciting feedback from a considerable number of real users, small-scale clinical trials if possible, and analysis of interaction data for continued system development. These pieces of feedback will provide valuable information to enhance the performance of the chatbot so that it may indeed be a useful and reliable tool in contributing to community mental health.

REFERENCES

[1] W. H. O. (WHO), "Depressive disorder (depression)," 31 March 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/depression. [Accessed 30 5 2025].

[2] A. D. M. J. D. J. A. S. C. C. M. A. Graves JM, "Geographic Disparities in the Availability of Mental Health Services in U.S. Public Schools," American Journal of Preventive Medicine, vol. 64, no. 1, pp. 1-8, 2023.

[3] L. S. C. S. Q. L. C. Y. S. M. L. X. Luo BA, "Mental health resources and its equity in Central South of China: A case study of Hunan Province," PLoS ONE, vol. 17, no. 10, p. e0272073, 2022.

[4] M. N. W. E. K. W. B. J. V. R. Pawaskar R, "Staff perceptions of the management of mental health presentations to the emergency department of a rural Australian hospital: qualitative study," BMC Health Services Research, vol. 22, no. 1, p. Article 87, 2022.

[5] K. C. R. T. S. K. G.-P. J. K. B. K. M. M. C. R. R. H. J. A. H. A.-M. J. D. K. G. W. G. O. K. S. M. K. M.-M. M. P. K. R. M. S. P. S. D. S. Montoya MI, "An international survey examining the impact of the COVID-19 pandemic on telehealth use among mental health professionals," Journal of Psychiatric Research, vol. 148, p. 188–196, 2022.

[6] W. H. O. (WHO), "Mental Health ATLAS 2020," 2021 .

[7] K. T. W. C. Kellie Yu Hui Sim, "Envisioning an AI-Enhanced Mental Health Ecosystem," in CHI '25 Workshop on Envisioning the Future of Interactive Health, 2025.

[8] N. R. H. H. E. W. S. E. S. J. V. J. M. A. C. P. &. R. Z. Eliane M. Boucher, "Artificially intelligent chatbots in digital mental health interventions: a review," Expert Review of Medical Devices, Vols. 18, Supplement 1, p. 37–49, 2021.

[9] B. K. Ahmedani, “Mental health stigma: Society, individuals, and the profession,” J. Soc. Work Values Ethics, vol. 8, no. 2, pp. 41–416, 2025, doi: 10.1177/22211117.

[10] S. Carolan and R. de Visser, “Employees’ perspectives on the facilitators and barriers to engaging with digital mental health interventions in the workplace: Qualitative study,” JMIR Ment. Health, vol. 5, no. 1, p. e8, 2018, doi: 10.2196/mental.9146.

[11] M. Strand, L. S. Eng, and D. Gammon, “Combining online and offline peer support groups in community mental health care settings: A qualitative study of service users’ experiences,” Int. J. Ment. Health Syst., vol. 14, p. 39, 2020, doi: 10.1186/s13033-020-00370-x.

[12] B. N. Renn, T. J. Hoeft, H. S. Lee, A. M. Bauer, and P. A. Areán, “Preference for in-person psychotherapy versus digital psychotherapy options for depression: Survey of adults in the U.S.,” NPJ Digit. Med., vol. 2, p. 6, Feb. 2019, doi: 10.1038/s41746-019-0077-1.

[13] I. Hungerbuehler, K. Daley, K. Cavanagh, H. Garcia Claro, and M. Kapps, “Chatbot-based assessment of employees’ mental health: Design process and pilot implementation,” JMIR Form. Res., vol. 5, no. 4, p. e21678, Apr. 2021, doi: 10.2196/21678.

[14] M. De Choudhury and S. De, “Mental health discourse on reddit: Self-disclosure, social support, and anonymity,” in Proc. Int. AAAI Conf. Weblogs Soc. Media (ICWSM), vol. 8, no. 1, pp. 71–80, May 2014.

[15] R. B. Jones and E. J. Ashurst, “Online anonymous discussion between service users and health professionals to ascertain stakeholder concerns in using e-health services in mental health,” Health Inform. J., vol. 19, no. 4, pp. 281–299, Dec. 2013, doi: 10.1177/1460458212474908.

[16] J. Koh, G. Y. Q. Tng, and A. Hartanto, “Potential and pitfalls of mobile mental health apps in traditional treatment: An umbrella review,” J. Pers. Med., vol. 12, no. 9, p. 1376, Aug. 2022, doi: 10.3390/jpm12091376.

[17] A. Sharma, A. S. Miner, D. C. Atkins, and T. Althoff, “A computational approach to understanding empathy expressed in text-based mental health support,” arXiv preprint, arXiv:2009.08441, 2020, doi: 10.48550/arXiv.2009.08441.

[18] B. O’Dell, K. Stevens, A. Tomlinson, et al., “Digital interventions in mental health: Evidence synthesis and future directions,” Evid. Based Ment. Health, vol. 25, pp. 45–46, 2022.

[19] J. Shen, D. DiPaola, S. Ali, M. Sap, H. Park, and C. Breazeal, “Empathy toward artificial intelligence versus human experiences and the role of transparency in mental health and social support chatbot design: Comparative study,” JMIR Ment. Health, vol. 11, p. e62679, 2024, doi: 10.2196/62679.

[20] X. Cheng, L. Su, and B. Yang, “An investigation into sharing economy enabled ridesharing drivers’ trust: A qualitative study,” Electron. Commerce Res. Appl., vol. 40, p. 100956, 2020, doi: 10.1016/j.elerap.2020.100956.

[21] H. Katoch, P. Jain, A. Sharma, Y. Sharma, and L. Gautam, “From algorithms to empathy: Navigating ethics, efficacy and user trust,” J. Adv. Res., vol. 2584-0142, 2025.

[22] A. Monn, T. V. de Araujo, A. Rüesch, G. Kronenberg, C. Hörmann, A. Adank, Z. Roman, G. Schoretsanitis, M. Rufer, E. Seifritz, and B. Kleim, “Randomized controlled trial for the Attempted Suicide Short Intervention Program (ASSIP): An independent non-replication study,” J. Affect. Disord., vol. 382, pp. 59–67, 2025.

[23] R. Dehbozorgi, S. Zangeneh, E. Khooshab, et al., “The application of artificial intelligence in the field of mental health: A systematic review,” BMC Psychiatry, vol. 25, p. 132, 2025, doi: 10.1186/s12888-025-06483-2.

[24] A. Thakkar, A. Gupta, and A. De Sousa, “Artificial intelligence in positive mental health: A narrative review,” Front. Digit. Health, vol. 6, p. 1280235, Mar. 2024, doi: 10.3389/fdgth.2024.1280235.

[25] R. Tornero-Costa, A. Martinez-Millana, N. Azzopardi-Muscat, L. Lazeri, V. Traver, and D. Novillo-Ortiz, “Methodological and quality flaws in the use of artificial intelligence in mental health research: Systematic review,” JMIR Ment. Health, vol. 10, p. e42045, 2023, doi: 10.2196/42045.

[26] J. Lyons-Cunha, “AI in mental healthcare: How is it used and what are the risks?,” Built In, Dec. 19, 2024. [Online]. Available: https://builtin.com/artificial-intelligence/ai-mentalhealth.

[27] World Health Organization, “Artificial intelligence in mental health research: New WHO study on applications and challenges,” WHO Europe News, Feb. 6, 2023. [Online]. Available: https://www.who.int/europe/news/item/06-02-2023-artificial-intelligence-in-mental-health-research--new-who-study-on-applications-and-challenges.

[28] American Psychological Association, “Artificial intelligence in mental health care,” APA Practice, Mar. 12, 2025. [Online]. Available: https://www.apa.org/practice/artificial-intelligence-mental-health-care.

[29] D. B. Olawade, O. Z. Wada, A. Odetayo, A. C. David-Olawade, F. Asaolu, and J. Eberhardt, “Enhancing mental health with artificial intelligence: Current trends and future prospects,” J. Med. Surg. Public Health, vol. 3, p. 100099, 2024, doi: 10.1016/j.glmedi.2024.100099.

[30] Q. Guo, J. Tang, W. Sun, H. Tang, Y. Shang, and W. Wang, “SouLLMate: An application enhancing diverse mental health support with adaptive LLMs, prompt engineering, and RAG techniques,” arXiv preprint, arXiv:2410.16322, 2024, doi: 10.48550/arXiv.2410.16322.

[31] 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, p. 31, Mar. 2024, doi: 10.1145/3643540.

[32] H. Lawrence, R. Schneider, S. Rubin, M. Matarić, M. McDuff, and M. Jones Bell, “The opportunities and risks of large language models in mental health,” JMIR Ment. Health, vol. 11, p. e59479, 2024, doi: 10.2196/59479.

[33] L. Boggavarapu, V. Srivastava, A. M. Varanasi, Y. Lu, and R. Bhaumik, “Evaluating enhanced LLMs for precise mental health diagnosis from clinical notes,” medRxiv preprint, 2024, doi: 10.1101/2024.12.16.24317648.

[34] “Emerging technologies (cs.ET); Neural and evolutionary computing (cs.NE); Optics (physics.optics),” arXiv preprint, arXiv:2410.01313, 2024, doi: 10.48550/arXiv.2410.01313.

[35] “OnRL-RAG: Real-time personalized mental health dialogue system,” arXiv preprint, arXiv:2504.02894, Apr. 2025, doi: 10.48550/arXiv.2504.02894.OnRL-RAG: Real-Time Personalized Mental Health Dialogue System. arXiv.org perpetual non-exclusive license

[36] “Comparing the best LLMs of 2025: GPT, DeepSeek, Claude & more – Which AI model wins?,” Sokada, 2025. [Online]. Available: https://www.sokada.co.uk/blog/comparing-the-best-llms-of-2025/.

[37] “The ultimate guide to the latest LLMs: A detailed comparison for 2025,” Empler, 2025. [Online]. Available: https://www.empler.ai/blog/the-ultimate-guide-to-the-latest-llms-a-detailed-comparison-for-2025.

[38] M. Casu, S. Triscari, S. Battiato, L. Guarnera, and P. Caponnetto, “AI chatbots for mental health: A scoping review of effectiveness, feasibility, and applications,” Appl. Sci., vol. 14, no. 13, p. 5889, 2024, doi: 10.3390/app14135889.

[39] L. Wang, T. Bhanushali, Z. Huang, J. Yang, S. Badami, and L. Hightow-Weidman, “Evaluating generative AI in mental health: Systematic review of capabilities and limitations,” JMIR Ment. Health, vol. 12, p. e70014, 2025, doi: 10.2196/70014.

[40] J. Swacha and M. Gracel, “Retrieval-augmented generation (RAG) chatbots for education: A survey of applications,” Appl. Sci., vol. 15, no. 8, p. 4234, 2025, doi: 10.3390/app15084234.

[41]A. N. Vaidyam, H. Wisniewski, J. D. Halamka, M. S. Kashavan, and J. B. Torous, “Chatbots and conversational agents in mental health: A review of the psychiatric landscape,”