Universidade Federal de São Carlos– UFSCar
Centro de Ciências Exatas e de Tecnologia– CCET
Departamento de Computação– DC
Programa de Pós-Graduação em Ciência da Computação– PPGCC

Rafael Vinicius Polato Passador

Personalizing Mental Health Support: A Retrieval-Based LLM Approach to Conversational Agent Development

São Carlos
2026

Rafael Vinicius Polato Passador

Personalizing Mental Health Support: A Retrieval-Based LLM Approach to Conversational Agent Development

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação do Centro de Ciências Exatas e de Tecnologia da Universidade Federal de São Carlos, como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação.

Área de concentração: Metodologias e Técnicas de Computação

Orientador: Helena de Medeiros Caseli

São Carlos
2026

Passador, Rafael Vinicius Polato

Personalizing Mental Health Support: A Retrieval-Based LLM Approach to Conversational Agent Development / Rafael Vinicius Polato Passador -- 2026.
108f.

Dissertação (Mestrado) - Universidade Federal de São Carlos, campus São Carlos, São Carlos
Orientador (a): Helena de Medeiros Caseli
Banca Examinadora: Altigran Soares da Silva, Helena de Medeiros Caseli, João Paulo Papa
Bibliografia

1. Inteligência Artificial. 2. Agente Conversacional. 3. Processamento de Linguagem Natural. I. Passador, Rafael Vinicius Polato. II. Título.

Ficha catalográfica desenvolvida pela Secretaria Geral de Informática (SIn)

DADOS FORNECIDOS PELO AUTOR

Bibliotecário responsável: Arildo Martins - CRB/8 7180

Folha de Aprovação

Defesa de dissertação de mestrado do(a) candidato(a) Rafael Vinicius Polato Passador, realizada em 24/04/2026

Comissão Julgadora

Prof(a) Dr(a) Helena de Medeiros Caseli (UFSCar)

Prof(a) Dr(a) João Paulo Papa (UNESP)

Prof(a) Dr(a) Altigran Soares da Silva (UFAM)

O relatório de defesa assinado pelos membros da comissão Julgadora encontra-se arquivado junto ao Programa de Pós-Graduação em Ciência da Computação

Dedico este trabalho a toda minha família. Em especial, minha mãe a Profa. Dra. Edna Aparecida Polato Passador, por sempre me incentivar a persistir o conhecimento.

Agradecimentos

Agradeço primeiramente a Deus, cuja bênção esteve presente em toda a minha jornada.

Ao corpo docente do Departamento de Computação da Universidade Federal de São Carlos, manifesto meu reconhecimento pelo valioso conhecimento compartilhado em suas dedicadas aulas e pelo esforço em auxiliar todos os alunos. Em especial, a Profa. Dra. Helena de Medeiros Caseli, meu agradecimento pela orientação e apoio. À Universidade Federal de São Carlos, meu agradecimento pela excelência de seu programa e estrutura.

Este trabalho foi desenvolvido com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) – Código de Financiamento 001. O trabalho também se insere no escopo do projeto AIM-Health, apoiado pela Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), processo nº 2024/10233-7, e pelo UKRI/MRC, aos quais estendemos nossos agradecimentos.

‘“Devemos mudar nossa atitude tradicional em relação à construção de programas. Em vez de imaginar que nossa principal tarefa é instruir o computador sobre o que ele deve fazer, vamos imaginar que nossa principal tarefa é explicar a seres humanos o que queremos que o computador faça“.

(Donald E. Knuth)

Resumo

Transtornos de saúde mental constituem um grande desafio global, exigindo soluções de suporte que sejam acessíveis e escaláveis. Agentes conversacionais têm emergido como ferramentas promissoras nesse contexto, mas sistemas baseados exclusivamente em modelos de linguagem de grande porte (LLMs) permanecem suscetíveis a saídas não fundamentadas, baixa ancoragem em conhecimento confiável e limitada adaptação ao contexto do usuário. Este trabalho investiga se a combinação de adaptação de domínio supervisionada, recuperação baseada em diretrizes e condicionamento por personas melhora a qualidade das respostas de suporte emocional. Três variantes, construídas sobre o mesmo backbone `Llama-3` ajustado por fine-tuning, são comparadas: um modelo Baseline com prompting em camadas, uma variante Hybrid Retrieval-Augmented Generation (RAG) ancorada em diretrizes de saúde mental da WHO e NICE por meio de recuperação híbrida, e uma variante Persona-Based RAG que incorpora sinais de persona tanto na recuperação quanto na construção do prompt. Todas as variantes são protegidas por um guardrail pré-LLM que intercepta entradas de alto risco e retorna uma resposta segura predefinida. A avaliação foi conduzida em um cenário de turno único, com entradas de datasets de saúde mental associadas a múltiplos perfis de persona. As saídas foram avaliadas por julgamento humano e por um protocolo de LLM-as-a-judge, sob um rubric comum que considera empatia, adequação ao tema e personalização. Os resultados convergem ao identificar a variante Persona-Based RAG como a melhor configuração geral, com ganhos mais claros em empatia e personalização, enquanto o RAG sem personas permaneceu competitivo em adequação ao tema. Esses achados indicam que a combinação de recuperação fundamentada em conhecimento com condicionamento por personas é uma estratégia promissora para o desenvolvimento de agentes conversacionais mais contextualmente adequados para suporte em saúde mental.

Palavras-chave: Agente Conversacional. Modelos de Linguagem. Personalização. Saúde Mental.

Abstract

Mental health disorders constitute a major global challenge, requiring support solutions that are both accessible and scalable. Conversational agents have emerged as promising tools in this context, but purely generative large language model (LLM) systems remain susceptible to unsupported outputs, weak grounding, and limited adaptation to user-specific context. This work investigates whether combining supervised domain adaptation, guideline-grounded retrieval, and persona-based conditioning improves the quality of emotional-support responses. Three system variants built on the same fine-tuned `Llama-3` backbone are compared: a Baseline model controlled through layered prompting, a Hybrid Retrieval-Augmented Generation (RAG) variant grounded in authoritative WHO and NICE mental health guidelines through hybrid retrieval, and a Persona-Based RAG variant that further incorporates persona signs into both retrieval and prompt construction. All variants are protected by a pre-LLM crisis guardrail that short-circuits high-risk inputs and returns a predefined safe response. Evaluation was conducted in a single-turn setting using inputs from mental health datasets associated with multiple persona profiles, and outputs were assessed through both human evaluation and an LLM-as-a-judge protocol under a shared rubric covering empathy, topic adequacy, and personalization. Results from both evaluation sources converge in identifying the Persona-Based RAG variant as the best overall configuration, with the clearest gains in empathy and personalization, while the RAG-only variant remained especially competitive in topic adequacy. These findings indicate that combining grounded retrieval with explicit persona conditioning is a promising strategy for the development of more contextually appropriate and supportive conversational agents for mental health.

Keywords: Conversational agent. Large Language Models. Personalization. Mental health.

List of Figures

List of Tables

Glossary

AI Artificial Intelligence

CA Conversational Agent

LLM Large Language Model

LLMs Large Language Models

CA Conversational Agents

NLP Natural Language Processing

RAG Retrieval Augmented Generation

Contents

Chapter 1

Introduction

1.1 Context

The approach and care of psychological disorders, especially depression and anxiety, which are among the most frequent ones, are currently seen as some of the main concerns in mental health. More recent official data from the Brazilian Ministry of Health indicate that 14.5% of adults aged 18 years or older living in the 26 state capitals and the Federal District reported a medical diagnosis of depression in 2024.[1]

In this scenario, the use of chatbots in mental health has seen a surge in interest due to their potential to bridge accessibility gaps and provide continuous support for individuals with mental health conditions (YU; MCGUINNESS, 2024; QIU et al., 2024; FAN et al., 2024). The widespread use of chatbots as conversational agents (CAs) in mental health can be attributed to their accessibility, cost-effectiveness, and ability to provide a nonjudgmental space for individuals to express themselves. According to Cho et al. (2023), these agents play a crucial role in addressing the treatment gap in mental health by making support more readily available, especially to individuals who may face barriers in accessing traditional therapy. Chatbots such as Woebot[2] and Wysa[3] have shown promise by offering continuous support, helping users manage mental health symptoms, and reducing stigma through private, convenient interactions.

Rule-based chatbots operate using pre-defined sets of rules and decision trees to guide user interactions (VAKAYIL et al., 2024). These systems excel in performing well-

[1] Brazilian Ministry of Health. *Vigitel Brasil 2006–2024*. Brasília: Ministry of Health, 2025. Available at: <https://www.gov.br/saude/pt-br/centrais-de-conteudo/publicacoes/svsa/vigitel/vigitel-brasil-2006-2024.pdf>. Accessed on: Mar. 31, 2026.

[2] Available at: <https://woebothealth.com/>

[3] Available at: <https://www.wysa.com/>

structured, task-specific dialogues where user inputs follow predictable patterns. As observed in Figure 1, this methodology is based on predefined responses that are triggered by the identification of keywords used by the user. However, due to their rigidity limits adaptability they struggle to handle varied expressions of the same intent, often failing to answer user queries outside their scripted paths. This framework suits scenarios where clear, stepwise guidance is required, such as providing static information or following strict workflows (MAENG; LEE, 2021). While reliable in their domains, rule-based systems lack contextual understanding, making them less effective in nuanced or emotionally sensitive applications like mental health support.

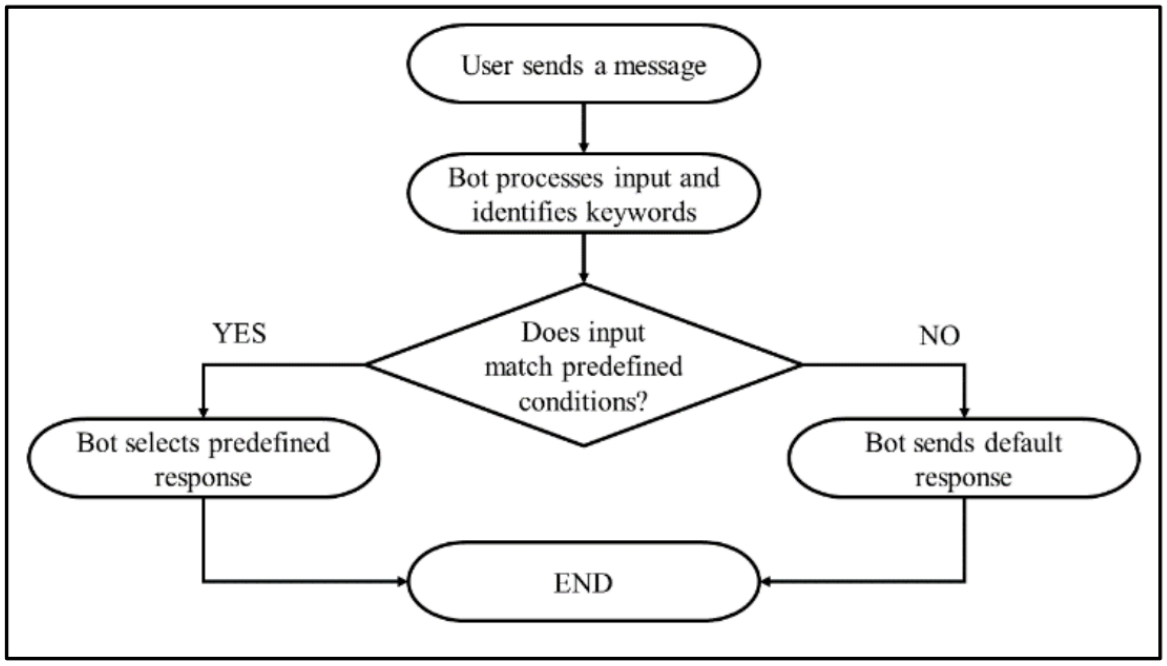

Figure 1 – Architecture of a rule-based ChatBot (VAKAYIL et al., 2024).

Retrieval-based CAs operate by selecting the most appropriate response from a predefined set of possible answers, rather than generating new responses, as shown in Figure 2. This approach is particularly useful in mental health applications because it enables the CA to provide accurate, controlled, and safe responses, crucial in contexts requiring reliability and empathy. Retrieval-based models employ information retrieval techniques, often based on similarity matching, to identify responses that best match a user's query. According to Cho et al. (2023), retrieval-based methods are frequently used in the medical domain as they offer a lower risk of "hallucination" (i.e., producing misleading or incorrect information), which is a common challenge in generative models.

Machine learning-based chatbots, often referred to as end-to-end chatbots, leverage data-driven models to generate responses dynamically. These systems utilize advanced techniques, like neural networks, to interpret user intent and provide contextually relevant, flexible answers. Large Language Models (LLMs) such as Llama-2 exemplify this approach, offering a high degree of conversational adaptability and emotional resonance,

essential for domains like mental health. However, despite their strengths, LLMs face significant challenges stemming mainly from the knowledge gaps within the models, including a lack of control over responses when the context is vague, potential hallucinations, limited contextualization within specific domains like mental health, and insufficient personalization for individual users (AGRAWAL et al., 2024).

As noted by Bora and CuayÁhuitl (2024), chatbots powered by Retrieval Augmented Generation RAG and LLM represent a promising advancement in healthcare. RAG systems enhance chatbot responses by integrating real-time retrieval of external data, allowing chatbots to generate responses that are not only contextually relevant but also factually grounded. This approach is particularly advantageous in the medical domain, where the accuracy and reliability of information are critical. RAG-based LLMs surpass traditional systems, enabling more human-like interactions while addressing the limitations of isolated rule-based or ML-based chatbots.

The Figure 2 shows an standard RAG system. Initially, a knowledge base (1) is processed by an embeddings model (2), which transforms textual data into numerical vector representations. When a user submits a query (3), it is processed through the same embeddings model (4) to generate a query embedding. This embedding is then used to search a vector store (5), where relevant documents or data points are retrieved based on similarity. The retrieved knowledge (6) is then sent to the LLM, which integrates the external data with its pre-trained knowledge to generate a final response (7) that is more informed and contextually relevant. This hybrid approach allows a chatbot to deliverer personalized and knowledge-rich responses, particularly beneficial for applications such as mental health support, where contextual understanding is crucial.

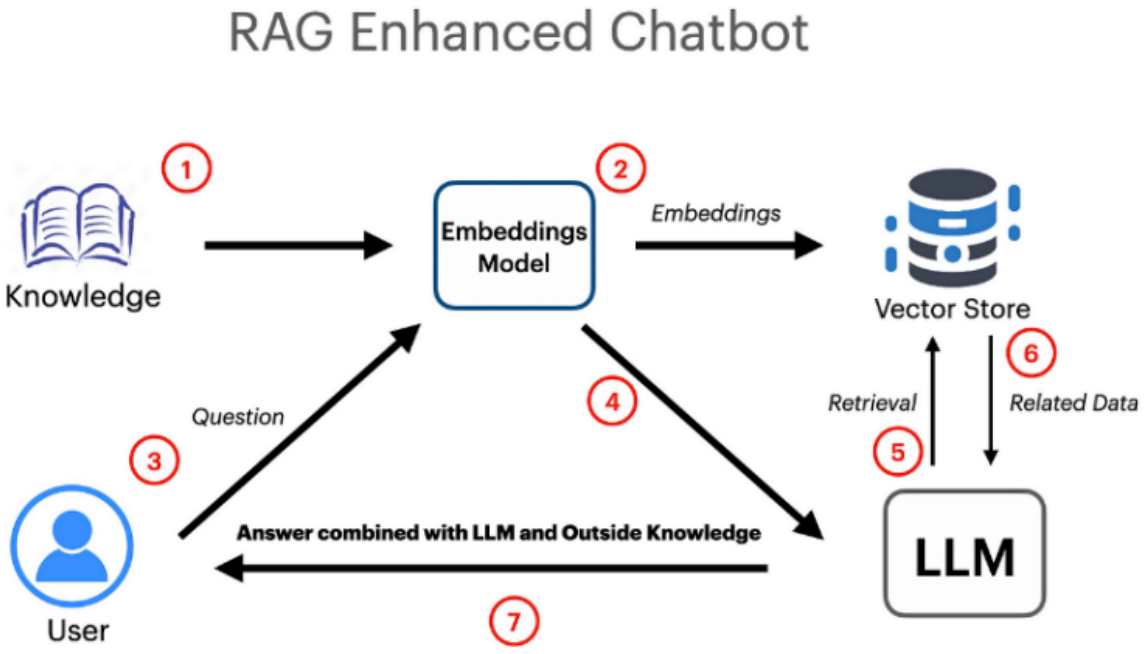

Figure 2 – Architecture of a RAG -based LLM ChatBot (BENITA et al., 2024).

These CAs maintain the advantage of stability and predictability, making them favorable for sensitive tasks, such as supporting mental health. However, one limitation is their reduced flexibility, as they rely on a fixed response database, which can limit personalization and adaptation to evolving conversational contexts.

LLMs such as Flan-T5, LLaMA-2, and GPT-like, when combined with RAG, have demonstrated improvements in handling complex queries (QIU et al., 2024; YANG et al., 2024; LI et al., 2024), thus proving their utility in medical chatbot applications. Instead of relying solely on pre-trained knowledge, retrieving the most relevant and up-to-date information for each user interaction and dynamically accessing external knowledge sources, significantly enhances the chatbot's ability to provide precise, context-aware responses, especially in specialized domains like mental health, where access to expert-curated information and individualized care is essential. This setup, when applied to mental health, can enable chatbots to provide informed, empathetic, and supportive interactions, tailored to the user's context and needs (BORA; CUAYÁHUITL, 2024).

Recent studies emphasize the importance of context-aware and user-adaptive dialogue systems in fostering meaningful engagement with individuals seeking psychological support (FAN et al., 2024; SANNA et al., 2024; YU; MCGUINNESS, 2024). Those authors elucidates that RAG-enhaced chatbots tailors responses based on prior interactions, user preferences, and clinically relevant knowledge, leading to more engaging and supportive conversation models. However, for RAG to effectively retrieve pertinent information, it is essential to have a well-structured knowledge base containing high-quality, domain-specific data. This includes curated mental health resources, user interaction histories, and expert-validated therapeutic content, which serve as the foundation for relevant and contextually appropriate retrieval. Without a sufficiently rich and representative dataset, the chatbot may struggle to provide responses that are both accurate and personalized, limiting its effectiveness in delivering meaningful support.

It is reasonable to assert that data resources are scarcer when working with non-English languages (JAFARI et al., 2025), posing a significant challenge for developing AI-driven mental health solutions. Motivated by the practical trade-offs reported in the literature for mental health NLP, this research investigates strategies that improve response assertiveness and personalization in domain-specific mental health tasks. In particular, prior evidence highlights that fine-tuning can yield stronger task performance but requires additional computational resources, whereas prompt engineering and RAG provide more flexible alternatives whose effectiveness depends on prompt design and retrieval quality (KERMANI; PEREZ-ROSAS; METSIS, 2025). Accordingly, this study evaluates a pipeline that combines supervised adaptation using the ESConv dataset Liu et al. (2021a), retrieval grounding with guideline documents, and persona conditioning as a central mechanism for personalization.

Despite data scarcity in Portuguese, the AMIVE dataset (ALVES et al., 2023), de-

veloped in a study focusing on college students' mental health, provides a rich source of contextualized user information that can serve as a foundation for personalization signals, including self-reported emotional states, responses to validated psychological scales (e.g., PHQ-9, WHODAS 2.0), behavioral indicators from wearables (heart rate, sleep quality, physical activity), and social media posts. In this thesis, personas and metadata are derived from AMIVE profiles and complemented with persona schemas from PersonaLens Zhao et al. (2025), enabling an explicit comparison of how different profile types affect personalization. Persona information is incorporated at two points of the response pipeline: it is injected into the prompt as contextual guidance for generation, and it is used to reformulate or enrich the retrieval query, so that evidence selection better matches the user context. This design is aligned with personalized RAG approaches, which argue that user-centric signals can improve both retrieval relevance and answer tailoring when integrated into retrieval and generation mechanisms (ZERHOUDI; GRANITZER, 2026a). Furthermore, Souza et al. (2022) provide design recommendations for chatbots supporting people with depression (including conversation style and personalization considerations), which can be extended to other mental health conditions and inform this study.

Therefore, this study seeks to highlight the potential of combining instruction-tuned LLMs, supervised domain adaptation with fine-tuning, retrieval grounding, and persona-based context engineering to produce responses that are simultaneously more assertive to the user's topic and more personalized to user context.

In this sense, the contribution of this study lies not only in the reported experimental results, but also in the methodological framework for investigating how fine-tuning, RAG, and persona-conditioned retrieval and prompting interact to shape response quality in specialized mental health scenarios.

This dissertation investigates whether the combination of supervised domain adaptation through LLM fine-tuning, guideline-grounded retrieval, and persona-based context engineering can improve the quality of emotional-support conversational agents for mental health, particularly with respect to topic adequacy and personalization. To this end, supervised fine-tuning was performed on ESConv (LIU et al., 2021a), a benchmark dataset for emotional support dialogue, and response generation was grounded in a knowledge base built from authoritative mental health guidelines published by the WHO[4] and NICE[5]. Retrieval was implemented through a hybrid search strategy, combining lexical and semantic signals in order to better support the grounding of generated responses in domain-relevant evidence.

In addition to grounding, this research investigates personalization as a complementary mechanism for improving emotional-support generation. Personalization is operationalized through explicit personas and structured metadata, allowing the system to adapt

[4] World Health Organization (WHO): <https://www.who.int/>
[5] National Institute for Health and Care Excellence (NICE): <https://www.nice.org.uk/>

both evidence retrieval and response generation to user-related context. Two persona sources are considered in order to analyze the effect of different profile schemas: AMIVE personas (SOUZA et al., 2025), which are centered on mental-health-related contextual attributes, and PersonaLens personas (ZHAO et al., 2025), which emphasize preference- and context-oriented traits for personalization evaluation. Persona signals are incorporated at two stages of the pipeline: first, by enriching the retrieval query to bias evidence selection toward user-relevant guidance; and second, by being injected into the prompt so as to guide generation toward more contextually appropriate responses. This design is also motivated by prior work showing that prompt structure and instruction sensitivity can substantially affect LLM behavior, thus reinforcing the importance of combining parameter adaptation with explicit prompting strategies (ZERHOUDI; GRANITZER, 2026b; TOPAL; BOZANTA; BASAR, 2024; CHATTERJEE et al., 2025).

Based on this design, the dissertation compares three system variants under the same task and evaluation rubric: a Baseline variant, consisting of a fine-tuned Llama-3 model controlled through prompt engineering, including role constraints, a structured planning scaffold, and a linguistic style guide to enforce a supportive and non-clinical stance;[6] a Hybrid RAG variant, which augments the baseline with retrieved guideline excerpts; and a Persona-Based RAG variant, which further extends Hybrid RAG by incorporating persona conditioning both during retrieval and during prompt construction. In all configurations, a crisis-detection guardrail is applied before retrieval and generation, so that high-risk inputs are intercepted and mapped to a deterministic safe response without invoking the LLM. To assess the contribution of each mechanism individually and in combination, the systems are evaluated in a single-turn setting, which enables a feasible and comparable protocol across variants. The evaluation considers three criteria—empathy, topic adequacy, and personalization—defined in a human evaluation manual reviewed by specialists in human-computer interaction and mental health, and adopts a two-stage protocol combining human judgments and an LLM-as-a-judge signal used as a complementary source of analysis rather than as a substitute for expert assessment (SZYMANSKI et al., 2025).

1.2 Objective

This dissertation aims to investigate, design, implement, and evaluate a personalized conversational pipeline for mental health support that combines supervised domain adaptation of large language models, retrieval-augmented generation, and persona-based conditioning with context engineering. The central objective is to analyze whether these complementary mechanisms can improve the quality of emotional-support responses in

[6] Llama-3-8B-Instruct was selected due to its strong reported performance, publicly available weights, and suitability for memory-constrained fine-tuning (GRATTAFIORI et al., 2024; KERMANI; PEREZ-ROSAS; METSIS, 2025).

a domain-specific setting, with particular emphasis on three dimensions adopted in this study: empathy, topic adequacy, and personalization.

To achieve this objective, open-weight instruction-tuned large language models are adapted through supervised fine-tuning on the ESConv dataset (LIU et al., 2021a), a benchmark for emotional support dialogue, in order to specialize the response style for supportive interactions. In parallel, a retrieval component is used to ground generation in authoritative mental health guideline documents published by the WHO[7] and NICE[8] so as to improve factual grounding and reduce unsupported content. Personalization is incorporated through explicit personas and structured metadata derived from two complementary sources: AMIVE personas (SOUZA et al., 2025), which provide mental-health-centered contextual profiles, and PersonaLens (ZHAO et al., 2025) personas, which emphasize preference- and context-oriented traits. Persona information is incorporated at two stages of the response pipeline: as contextual guidance in prompt construction and as an enrichment signal for retrieval queries, so that evidence selection can be better aligned with user-specific context.

Overall, this work seeks to provide an empirical comparison of three system variants sharing the same fine-tuned backbone: a baseline model controlled through prompt engineering, a retrieval-augmented variant grounded in guideline documents, and a persona-based RAG variant that further integrates persona conditioning into retrieval and generation. To assess the contribution of each mechanism individually and in combination, the systems are evaluated under a controlled single-turn protocol, chosen to ensure feasibility and comparability across variants, using both specialist human evaluation and an LLM-as-a-judge procedure. Through this comparison, the dissertation aims to determine to what extent retrieval grounding and persona-based conditioning contribute to the development of more appropriate, context-sensitive, and supportive conversational agents for mental health.

1.3 Document Organization

This chapter has presented a brief overview, outlining the context, motivation, and overarching objective of this study. The remainder of this document is organized as follows:

- ❑ Chapter 2 presents an in-depth description of the theoretical background pertinent to this research.
- ❑ Chapter 3 reviews existing work and prior studies that are closely related to this investigation.

[7] Available at: <https://www.who.int/>
[8] Available at: <https://www.nice.org.uk/>

- Chapter 4 describes the resources employed in this study, including the datasets used for fine-tuning and evaluation, the document collection curated for the knowledge base, and the persona sources and metadata schemas adopted for personalization.
- Chapter 5 provides a comprehensive description of the employed methodology and the experimental setup, including all relevant implementation details, evaluation procedures, and parameters used throughout the study.
- Chapter 6 presents and discusses the experimental results, comparing the baseline, RAG, and PersonaRAG variants under both human evaluation and LLM-as-a-judge assessment, and analyzing the impact of retrieval grounding and persona conditioning across the evaluation criteria.
- Chapter 7 finishes this dissertation summarizing the main conclusions and limitations of this work but also pointing to future directions.

Chapter 2

Theoretical Foundation

This section provides the theoretical foundation necessary to understand the development of conversational agents, with a particular focus on chatbots leveraging LLMs and RAG. First, we introduce the concept of conversational agents, outlining the main techniques used in the literature for their development, including rule-based systems, machine learning-based approaches, and the integration of LLMs with RAG. Next, we offer a concise explanation of LLMs, detailing their architecture, training process, and general application. Subsequently, we discuss RAG, describing its role in enhancing LLM-based chatbots by incorporating external knowledge retrieval mechanisms. Finally, we present an overview of RAG-based chatbot evaluation methods, emphasizing key metrics and approaches used to assess their effectiveness in various contexts.

2.1 Chatbots as Conversational Agents

Conversational Agents (CA), also referred to as chatbots or dialogue systems, are systems designed to interact with users through natural language. The development and adoption of these agents have been significantly influenced by advances in Natural Language Processing (NLP), enabling them to understand, interpret, and generate human-like responses (CHO et al., 2023). Conversational Artificial Intelligence (AI), illustrated in Figure 3, surround a broader range of capabilities that facilitate human-machine interaction, including speech recognition, sentiment analysis, and context-aware responses. Despite these advancements, fully achieving near-human conversational abilities remains a challenge due to the complexity of human cognition and linguistic nuances.

Chatbots can be classified based on their underlying approach to generating and selecting responses. The three main categories include rule-based chatbots, machine learning-

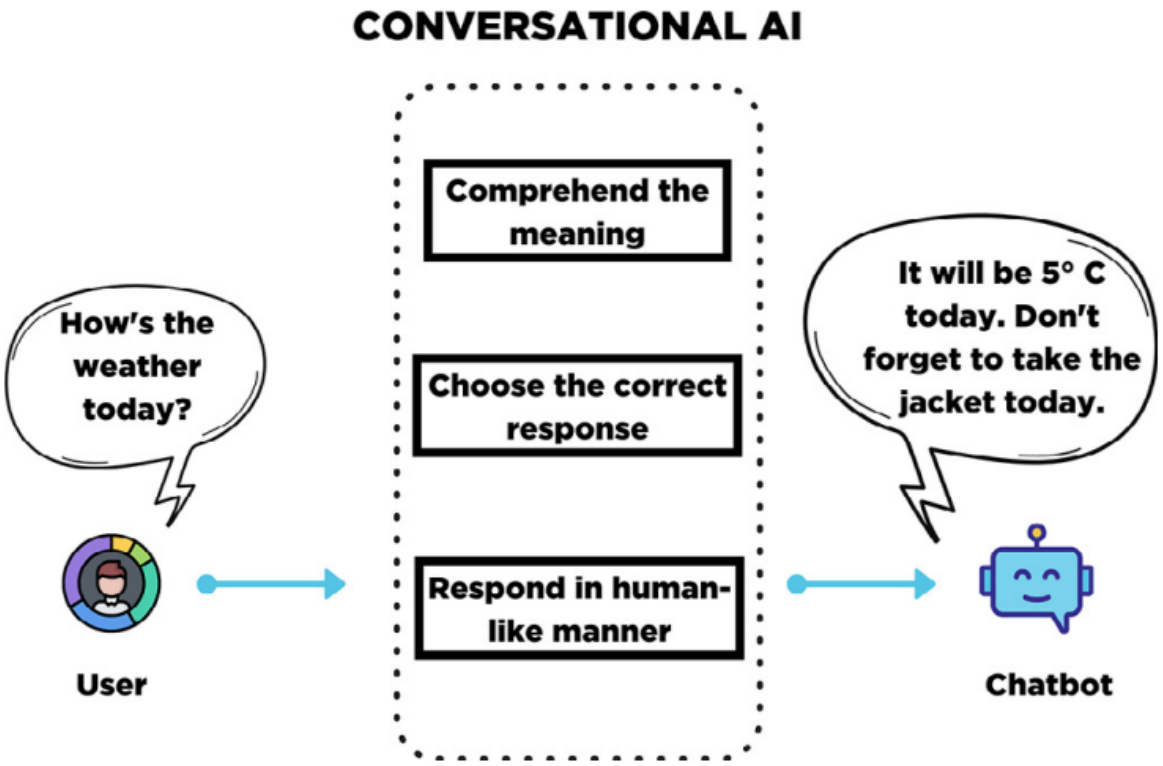

Figure 3 – Overview of a Conversational AI Chatbot (SINGH; BENIWAL, 2022).

based chatbots, and LLM+RAG-based chatbots, each differing in their adaptability, complexity, and effectiveness in various applications.

- **Rule-based Chatbots:** These chatbots operate on predefined sets of rules and decision trees, determining responses based on structured patterns and keyword recognition (VAKAYIL et al., 2024). They follow a deterministic approach, where each user input is mapped to a fixed response, making them suitable for simple and structured interactions such as booking appointments or providing frequently asked questions (FAQs). However, their rigidity and limited ability to handle unexpected user inputs can lead to frustrating user experiences, particularly in dynamic or complex conversations.

- **Machine Learning-based Chatbots:** These chatbots utilize deep learning models, such as neural networks, to analyze user inputs and generate contextually relevant responses (SINGH; BENIWAL, 2022). Unlike rule-based systems, they do not rely on predefined rules but instead learn from large datasets, enabling them to generalize across diverse conversational patterns. ML-based chatbots are particularly effective in scenarios requiring natural, fluid, and adaptive conversations. However, they may still be constrained by the quality and diversity of the training data, limiting their ability to address highly specific or unseen queries. Additionally, they lack the ability to retrieve external, real-time information beyond their training corpus, which can be a drawback in applications requiring up-to-date knowledge.

- **LLM+RAG-based Chatbots:** The integration of LLM and RAG represents a significant advancement in chatbot technology. LLMs, trained on vast amounts of text data, exhibit superior language comprehension and generation capabilities, allowing them to produce responses with greater coherence, fluency, and contextual awareness. However, their reliance on static training data limits their ability to retrieve and process real-time or domain-specific knowledge.

  The RAG framework addresses this limitation by incorporating an information retrieval mechanism, enabling chatbots to fetch relevant documents or knowledge sources dynamically before generating responses. This approach enhances the accuracy and reliability of chatbots in specialized domains, such as healthcare and legal assistance, where precise and evidence-backed responses are crucial (VAKAYIL et al., 2024). LLM+RAG-based chatbots provide a more personalized and informed conversational experience, making them particularly well-suited for applications requiring nuanced understanding and real-time knowledge retrieval.

Brazil leads the world in prevalence of anxiety disorders and ranks fifth in depression rates, as highlighted by (SOUZA; SOUSA, 2017). These statistics underscore the urgent need for scalable and accessible mental health interventions. In this context, Cho et al. (2023) emphasizes the growing attention towards the use of chatbots, in mental health care within both computer science and medical research domains. These systems offer a promising solution by providing a less intimidating, anonymous platform for mental health support, effectively bridging the gap between the demand for care and its limited availability.

The literature provides compelling evidence of the viability of mental health chatbots through examples like (FITZPATRICK; DARCY; VIERHILE, 2017) and (INKSTER; SARDA; SUBRAMANIAN, 2018), both of which have demonstrated effectiveness in delivering therapeutic interventions such as cognitive-behavioral therapy (CBT). Woebot, for instance, has been shown to reduce symptoms of depression and anxiety in young adults through interactive CBT sessions, while Wysa leverages empathy-driven interactions to support users in managing stress and emotional challenges. These systems underscore the potential of conversational agents to address accessibility barriers in mental health care, offering scalable and personalized support. However, significant gaps remain, particularly in addressing cultural and linguistic diversity, ensuring sustained user engagement over time, and integrating advanced technologies LLM to improve conversational depth and adaptability. Furthermore, many existing chatbots are limited in their ability to handle complex, multifaceted mental health conditions, highlighting the need for ongoing research to expand their capabilities and impact.

Over the past decade, extensive research has assessed the feasibility and efficacy of chatbots, particularly in improving mental health outcomes. Systematic reviews have

consistently highlighted their effectiveness in healthcare settings, as well as their suitability for supporting individuals with chronic conditions by promoting health behavior changes such as increased physical activity, improved diet, and weight management (CHAKRABORTY et al., 2023). He et al. (2022) tested the clinical effectiveness and nonclinical performance of a cognitive behavioral therapy (CBT)–based mental health chatbot (XiaoE) for young adults with depressive symptoms during the COVID-19 pandemic, and showed that the chatbot significantly reduced depressive symptoms in college students compared to control groups, with a moderate effect size post-intervention and a small effect size at 1-month follow-up. In parallel, Vereschagin et al. (2024) in a randomized trial of an AI chatbot delivering cognitive–behavioral therapy, reduced anxiety and depressive symptoms, improved mental well-being, and decreased the frequency of cannabis use and alcohol consumption among university students.

These studies and trials (CHAKRABORTY et al., 2023; HE et al., 2022; VERESCHAGIN et al., 2024) assures the viability of using AI chatbots in mental health domain.

2.2 Large Language Models

LLMs and their associated technologies are widely regarded as one of the most impactful advancements in recent AI technology. They are built upon Transformer-based architectures, which leverage self-attention mechanisms to model long-range dependencies within textual data (VASWANI et al., 2017). Their development has been driven by advancements in deep learning, self-supervised learning, and scaling laws, which emphasize increased model size and training data as key factors in improving performance. These models typically consist of billions of parameters and are trained on extensive text corpora sourced from publicly available data, web crawls, academic literature, and other structured knowledge bases. Some of the most well-known LLMs include GPT-4, Gemini, Claude, and Llama.

The evolution of LLMs has been marked by a shift from earlier rule-based and statistical models to neural architectures that employ Transformer-based learning. Unlike recurrent and convolutional networks, which have limitations in handling long-range dependencies, the Transformer architecture relies on a self-attention mechanism that dynamically focuses on different parts of the input text, enabling improved comprehension of syntactic and semantic relationships (VASWANI et al., 2017). These models are typically trained using autoregressive or masked language modeling techniques, which help them predict missing words or generate coherent responses in a conversational setting.

The training process of LLMs follows a multi-stage approach. In the pre-training phase, the model learns general linguistic patterns and world knowledge from vast corpora without explicit human supervision. After pre-training, the model undergoes fine-tuning, where it is adapted to specific tasks using labeled datasets. This step refines the

model's performance for particular applications such as legal document analysis, medical diagnosis, or chatbot-based mental health interventions. More recent LLMs incorporate reinforcement learning from human feedback, a method in which human annotators rank responses to guide the model toward generating more accurate, ethical, and contextually relevant outputs. The Figure 4 illustrates this training process. Llama 2-Chat, follows a structured pipeline that integrates self-supervised learning, supervised fine-tuning, and reinforcement learning with human feedback (RLHF) to enhance its conversational capabilities. This iterative refinement process allows the model to adapt dynamically based on human feedback, ensuring more coherent, reliable, and safe interactions.

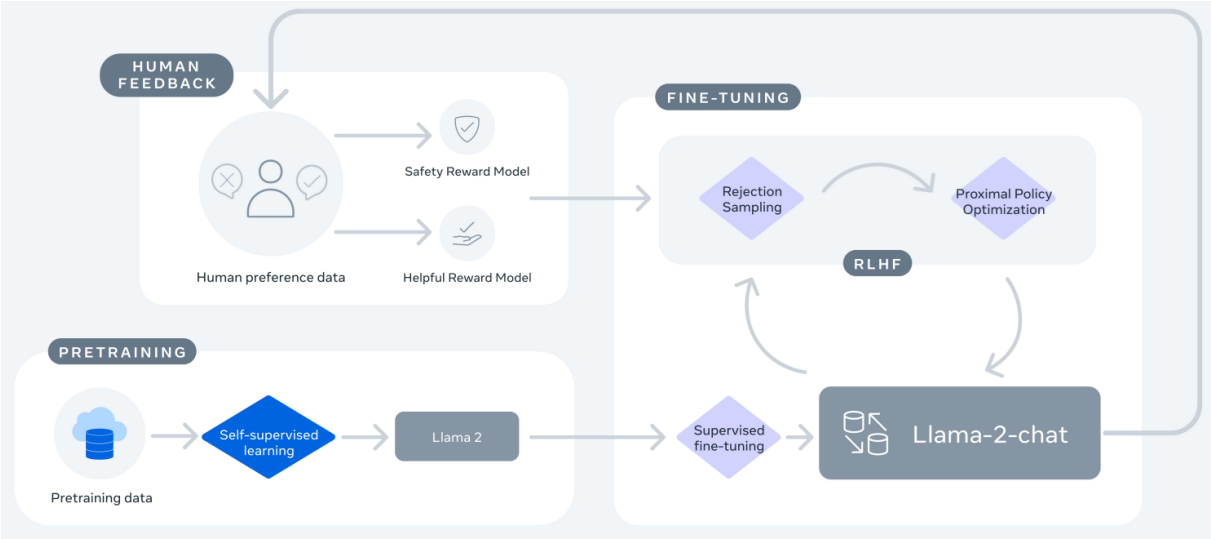

Figure 4 – Training of Llama 2-Chat (TOUVRON et al., 2023).

Beyond training and fine-tuning, prompting techniques enhance LLMs by guiding response generation, as shown by diverse studies in this area (LI et al., 2024; YANG et al., 2024; LI et al., 2024). Zero-shot prompting relies solely on pre-trained knowledge, while Few-shot prompting improves task understanding with a few examples. Chain-of-Thought (CoT) encourages step-by-step reasoning, and Tree-of-Thought (ToT) structures responses as decision trees, enhancing problem-solving. Role-playing prompts instruct the model to adopt specific personas, such as a therapist, improving contextual relevance and empathy. These strategies refine chatbot interactions, ensuring coherence, accuracy, and sensitivity in mental health applications.

Despite their strengths, LLMs face significant challenges, particularly in terms of hallucinations—where models generate incorrect or misleading information that appears plausible but lacks factual accuracy. This issue is especially critical in domain-specific applications, such as mental health support, where incorrect or non-contextualized responses can have serious consequences for users seeking guidance. The reliability of LLMs is further complicated by their inherent limitations in understanding nuanced human emotions, cultural contexts, and ethical considerations, which are essential for fostering meaningful and empathetic interactions. Additionally, biases embedded in training data can lead to re-

sponses that reinforce stereotypes or exclude marginalized perspectives, raising concerns about fairness and inclusivity. Addressing these challenges requires not only improved training methodologies, such as reinforcement learning with human feedback (RLHF), but also more sophisticated mechanisms for ensuring context information to guarantee high accuracy and conformability.

These limitations are particularly relevant for the experimental design adopted in this thesis. Rather than relying only on the pre-trained behavior of a general-purpose model, the final system combines supervised fine-tuning, instruction design, few-shot demonstrations, and persona conditioning to adapt the backbone model to the emotional-support domain in a more controlled and reproducible way.

2.3 Retrieval Augmented Generation (RAG)

Given the complexity of healthcare data and the need for precise information, the application of RAG-based LLMs represents a significant step forward in AI-assisted healthcare technologies (BORA; CUAYÁHUITL, 2024).

A primary issue of LLM is their tendency to produce "hallucinations", generating plausible yet nonfactual responses when confronted with queries that exceed their training data or demand up-to-date information (HUANG et al., 2024). To address these limitations, RAG improves LLM performance by retrieving relevant document segments from external knowledge bases using semantic similarity measures. By incorporating external references, RAG minimizes the risk of generating inaccurate content and has become an essential component in advancing chatbot technologies. Its integration has expanded the practical applications of LLMs, making them more effective for real-world scenarios

The retrieval process in the RAG framework plays a crucial role in enhancing the accuracy and contextual relevance of generated responses. As shown in Figure 5, once documents are indexed, they are split into smaller chunks, encoded into numerical vector representations, and stored in a vector database. During retrieval, when a user submits a query, the system computes the semantic similarity between the query and the pre-encoded document vectors. Using efficient similarity search algorithms, the model selects the top-k most relevant document chunks, ensuring that the retrieved information is contextually aligned with the user's request. These selected chunks are then passed to the generation stage, where the LLM combines the retrieved context with the query to produce a more informative and context-aware response. This retrieval mechanism is particularly beneficial in dynamic knowledge domains, where real-time, factual grounding is essential to minimize hallucinations and enhance reliability.

Different retrieval techniques are commonly used in the RAG framework to enhance information retrieval and provide relevant context to LLMs. BM25 is a ranking function based on term frequency-inverse document frequency (TF-IDF) that retrieves documents

by measuring the lexical similarity between a query and stored text. Keyword search operates by identifying exact word matches within documents, offering a straightforward and efficient method for retrieving relevant passages. More recent techniques, such as cross-encoders, process the query and document together through a deep learning model to generate a relevance score, enabling context-aware retrieval. Each of these methods plays a distinct role in optimizing retrieval, depending on the required balance between efficiency and semantic precision.

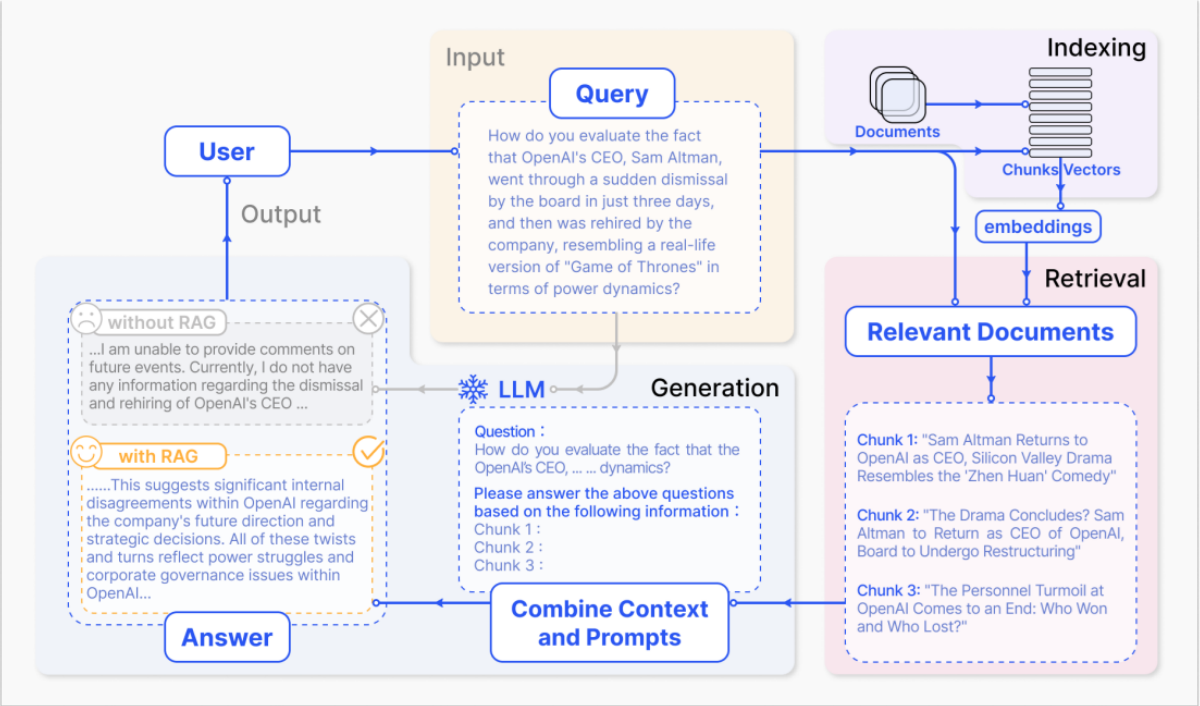

Figure 5 – A representative instance of the RAG (GAO et al., 2024).

In this context, many works already proven the capability of using RAG technics in LLM-based chatbots to provide specializaed context-knowledge to the model (VAKAYIL et al., 2024; YANG et al., 2024; SANNA et al., 2024; FAN et al., 2024; ABBASIAN et al., 2024). Furthermore, in pursuit of personalized user responses, Abbasian et al. (2024) employ a retrieval-based approach to access patient-specific information, maintaining a history of user interactions collected through question-answer exchanges and incorporating these details into the LLM prompt. Similarly, Yang et al. (2024) apply a comparable strategy in the domain of dietary personalization, leveraging users' eating habits to tailor nutritional recommendations effectively. Both approaches highlight the role of retrieval-augmented generation (RAG) in enhancing the contextual relevance and personalization of LLM outputs, demonstrating its applicability across diverse domains.

In this work, the general RAG principle is instantiated through a hybrid retrieval pipeline that combines lexical and dense retrieval signals, followed by a fusion stage, so that both term overlap and semantic similarity contribute to evidence selection. Moreover, retrieval is not treated here as a purely generic grounding mechanism: in the personalized

variant, persona metadata is used to reformulate the query so that retrieved passages better match the user context, and a crisis-detection guardrail is applied before retrieval to short-circuit high-risk cases with a deterministic response. These design choices refine the role of RAG from simple document injection to a controlled grounding layer that supports safety, relevance, and personalization simultaneously.

2.4 Evaluation

Evaluating LLMs not only helps us better understand the strengths and weakness of them, but also paramount importance of ensuring their safety and reliability, particularly in safety-sensitive sectors such as financial institutions and healthcare facilities (CHANG et al., 2024).

First, existing quantitative evaluation methods evaluate chatbots outcomes based on language-specific perspectives and surface-form similarity (QIU et al., 2024; DAS et al., 2022; LI et al., 2024; YU; MCGUINNESS, 2024), employing metrics such as Bilingual Evaluation Understudy (BLEU) (PAPINENI et al., 2002), Recall-oriented Understudy for Gisting Evaluation (ROUGE) (LIN, 2004), Perplexity (JELINEK et al., 2005), BERT-Score (**??**) and METEOR (BANERJEE; LAVIE, 2005).

BLEU, calculates n-gram precision, comparing how many words or phrases in the chatbot's response match the reference while applying a brevity penalty to prevent excessive short answers. ROUGE, on the other hand, commonly applied in text summarization, evaluates how much of the reference text is recovered in the chatbot's response. Key variations include ROUGE-N, which measures exact n-gram matches, ROUGE-L, which considers longest common subsequences, and ROUGE-S, which accounts for skip-grams that capture words appearing in different orders.

To overcome limitations of n-gram based metrics, such as failing to capture semantic equivalence or the emotional and empathetic quality of the response, BERTScore introduces a semantic similarity approach by comparing word embeddings instead of direct word matches. Based on transformer models like BERT, this metric assesses how similar two texts are in meaning, even if they use different words. This is particularly relevant for mental health chatbots, where responses may be paraphrased yet still meaningful and supportive.

METEOR (Metric for Evaluation of Translation with Explicit ORdering) improves upon BLEU by incorporating synonym matching, stemming, and word order considerations. This makes it more effective for evaluating open-ended, diverse responses in mental health conversations, as it accounts for linguistic variation and paraphrasing.

Recently, GPTScore (FU et al., 2024) has emerged as a deep-learning-based evaluation metric, where a pre-trained GPT model assigns a quality score to chatbot responses. Unlike traditional methods, GPTScore leverages contextual understanding, evaluating

responses based on coherence, factual correctness, empathy, and engagement. This approach is particularly well-suited for assessing depression-focused chatbots (FAN et al., 2024), as it can determine whether responses are emotionally supportive, non-triggering, and contextually relevant to the user's needs.

Although automatic evaluation metrics provide valuable insights into chatbot performance, human evaluation remains essential, particularly in sensitive domains such as mental health support. Those quantitative metrics focus on linguistic similarity and fluency but can fail to capture emotional appropriateness, contextual relevance, and empathetic engagement, which are critical for interactions with individuals experiencing depression. Furthermore, the scarcity of high-quality, domain-specific test datasets, makes it difficult to rely solely on automated metrics, as they may not adequately reflect real-world user experiences. Human annotators can assess responses based on coherence, factual accuracy, emotional support, and therapeutic effectiveness, ensuring that chatbots provide meaningful and responsible interventions.

In Qiu et al. (2024) professional counselors are presented with a dialogue history and three randomly shuffled responses (baseline, fine-tuned, ground truth). They are tasked with selecting the optimal response between every two responses for the dialogue history, considering aspects such as professionalism, informativeness, helpfulness, empathy, and safety. (FAN et al., 2024), on the other hand, used a scoring-based subjective evaluation method to understand the effects of model outputs, with independent evaluators scoring professionalism, fluency, and empathy.

Accordingly, the experiment conducted in this work does not rely on a single automatic metric to compare conversational variants. Automatic metrics are mainly used to support model selection and fine-tuning analysis, whereas the final comparison among the Baseline, RAG, and Persona-Based RAG systems is performed in a single-turn setting through human judgments over empathy, topic adequacy, and personalization, complemented by an LLM-as-a-judge stage. This evaluation protocol is also conducted in English, which is consistent with the language of the selected benchmark inputs and model prompts; Portuguese translations, when provided, serve only as comprehension support for specialist human judges.

Chapter 3

Related Works

In this chapter, we provide an overview of research related to our study, with a primary emphasis on works that address LLMs based ChatBots in Mental Health and others domains.

Recent advancements in Artificial Intelligence (AI) and Natural Language Processing have introduced the potential for large language models (LLMs) to support mental health services by providing real-time, empathic conversational responses. Thus, this chapter briefly describes some works that used LLMs to develop and evaluate chatbots in mental health or other domains.

3.1 Yu and McGuinness (2024)

In a study, Yu and McGuinness investigate a novel approach to developing a chatbot for mental health support that leverages the integration of fine-tuned LLMs with prompt-based augmentation. Their primary objective is to improve conversational quality, coherence, and user safety in mental health interactions, areas where traditional LLMs often fall short due to the complexities of emotional dialogue and the sensitive nature of mental health care.

The methodology employed by Yu and McGuinness (2024) centers on the development of a hybrid chatbot model that combines the strengths of two AI systems: a fine-tuned version of DialoGPT, optimized for therapy-oriented conversations, and ChatGPT 3.5, which supports real-time, broader contextual interaction. The research involved fine-tuning DialoGPT using a dataset comprising approximately 5,000 therapy-oriented conversations, structured to capture authentic therapeutic dialogue patterns. To achieve this,

the researchers sourced conversational therapy transcripts from publicly available data[1], ensuring that the language model would reflect a range of therapeutic interactions. Data preprocessing was applied to anonymize personal information, generalize specific details, and remove idiomatic expressions that might reduce the model's generalizability.

The architecture proposed is designed to integrate the domain-specific insights of the fine-tuned DialoGPT model with the flexible, real-time conversational abilities of ChatGPT, as shown in Figure 6. Specifically, DialoGPT generates initial responses that serve as "contextual knowledge injections," which are used to guide ChatGPT's subsequent output. In this arrangement, the ChatGPT API accesses both the original user input and the DialoGPT-generated prompt, allowing it to construct responses that are both contextually informed and linguistically sophisticated. This approach attempts to enhance the chatbot's overall empathy, responsiveness, and alignment with therapeutic standards, creating a conversational experience that can better support mental health needs.

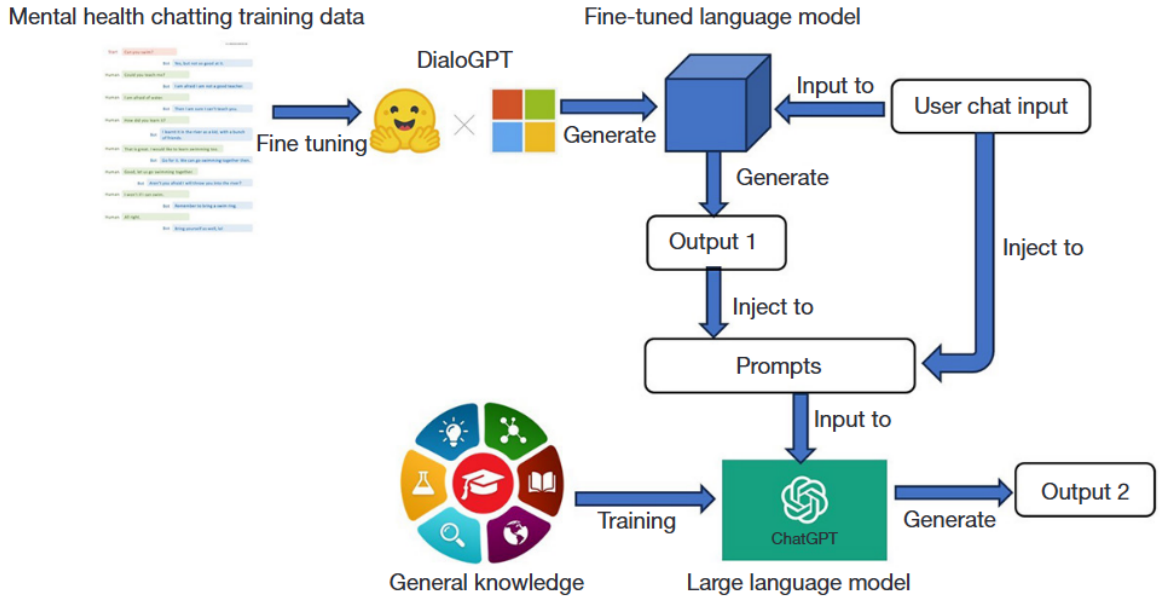

Figure 6 – Chatbot architecture proposed by Yu and McGuinness (2024).

Evaluation of the chatbot's effectiveness was performed using both quantitative and qualitative metrics. Automated assessments, including perplexity and BLEU scores, were used to evaluate the accuracy and fluency of generated responses. In these tests, the hybrid model showed a significant improvement over standalone ChatGPT and DialoGPT configurations, achieving lower perplexity and higher BLEU scores, which indicate enhanced conversational relevance and coherence. Additionally, the authors conducted human evaluations by soliciting feedback from two groups: users with mental health concerns and mental health professionals. Surveys from these groups revealed high satisfaction with the chatbot's language quality, empathetic responses, and overall conversational flow. Sugges-

[1] Available at: <http://www.thetherapist.com/Transcripts.html>

tions for improvement highlighted the need to increase response variety, avoid repetitive phrasing, and enhance emotional support through more nuanced language.

The results of the study demonstrate the potential of LLMs to contribute meaningfully to mental health support systems when integrated with domain-specific knowledge and prompt engineering. Both users and professionals indicated positive experiences with the chatbot's interactions, particularly regarding its sensitivity to emotional cues and capacity for providing supportive responses. However, the study also underscore critical limitations that require further research: the need to enhance user privacy, refine the model's ability to avoid potentially risky responses in sensitive contexts, and develop adaptive conversational techniques that respond to users' evolving emotional states.

In conclusion, the study contributes valuable insights to the development of LLM-driven mental health chatbots by demonstrating the efficacy of a hybrid architecture that leverages the strengths of both fine-tuned models and prompt engineering. While promising, their findings also emphasize the necessity of addressing ethical and practical challenges, particularly those related to privacy, safety, and contextual understanding, to enable responsible deployment of these tools in mental health care.

3.2 Vakayil et al. (2024)

In response to the growing need for supportive, accessible, and sensitive communication tools for individuals affected by sexual harassment, Vakayil et al. developed a chatbot leveraging LLMs to assist victims. Utilizing the Llama-2 model, this chatbot integrates RAG techniques to generate responses that are both contextually relevant and empathetic. The goal is to provide to users a conversational agent with accurate, nonjudgmental information and guidance in a compassionate manner, recognizing that individuals affected by sexual trauma often hesitate to disclose their experiences due to fear, shame, and the potential for societal stigma.

The methodology proposed by the study employs the Llama-2-7b model, a member of Meta's Llama-2 family which is built on a transformer architecture. This model uses a supervised, fine-tuning approach that is well-suited for handling complex dialogues. It provides responses by building on a vast, pre-trained corpus while adding task-specific knowledge through RAG techniques. In this study, the authors combined Llama-2 with a fine-tuned BERT model, which maps text data to a vector space, allowing the chatbot to identify and retrieve contextually appropriate information from a ChromaDB vector database.

The workflow developed begins with document parsing and text preprocessing, where relevant PDF files containing information on sexual harassment laws and resources in India are split into semantically meaningful chunks. Each chunk is then embedded using the fine-tuned BERT-base model to create vector representations that the chatbot can

later use to retrieve relevant information based on user queries. When a user inputs a query, the chatbot performs a vector search in the ChromaDB using cosine similarity to find the most relevant documents, which are then provided as context for Llama-2 to generate a human-understandable response. This RAG -based setup, as shown in Figure 7, allows the chatbot to combine retrieved, up-to-date knowledge with LLM generation, producing responses tailored to the user's specific situation and needs

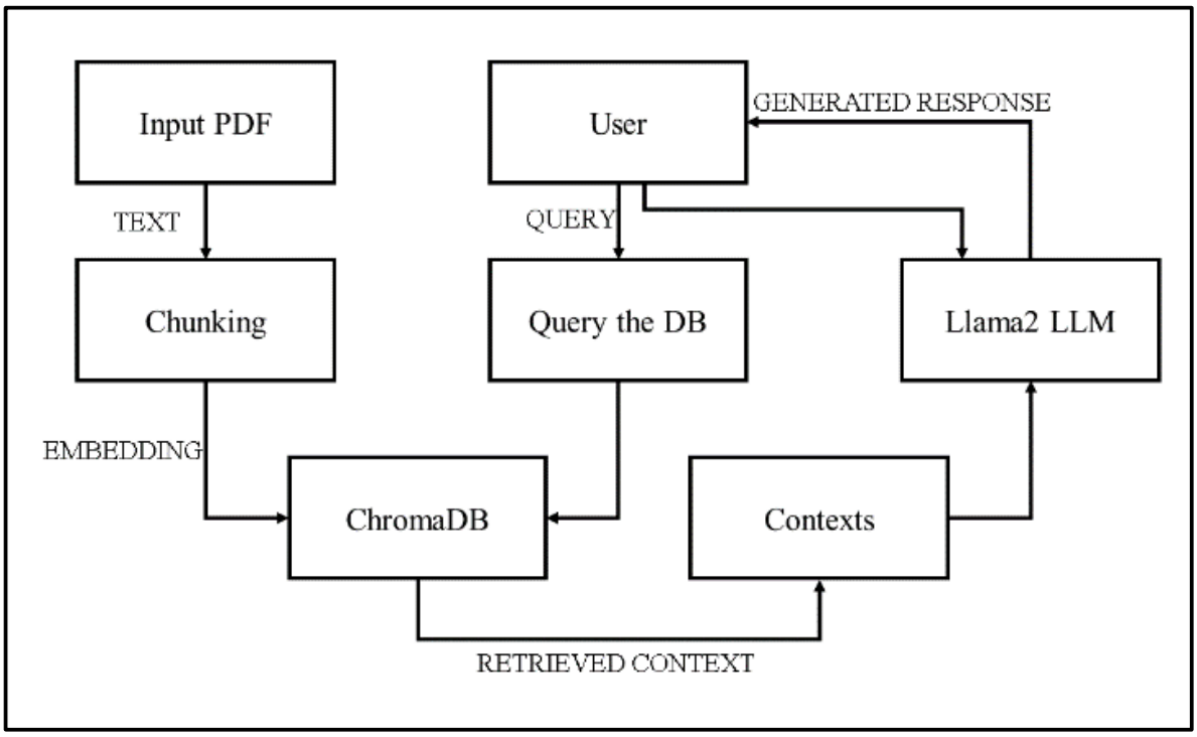

Figure 7 – Llama-2-7b RAG chatbot (VAKAYIL et al., 2024).

Throughout the study, the authors emphasized response quality and sensitivity, integrating response guidelines into the model's prompt design to prioritize empathy, safety, and nonjudgmental language. To accomplish this, the chatbot's prompt template instructs it to respond in an understanding, non-blaming tone, while always encouraging users to seek professional help as appropriate. By grounding its responses in this ethical framework, the chatbot aims to create a supportive environment for victims of harassment, addressing one of the significant barriers to seeking help.

In terms of evaluation, the chatbot demonstrated a high degree of effectiveness, with response accuracy exceeding 95%, indicating that it was generally able to retrieve and present relevant, factually accurate information. While the chatbot performed well in terms of accuracy, Vakayil et al. (2024) note occasional inconsistencies in responses, such as retrieving the wrong helpline for male victims of harassment, which they attribute to the limited availability of resources for men within the dataset. Despite these limitations, the chatbot consistently delivered empathetic and nonjudgmental responses that acknowledged the user's emotions, provided support, and included practical information

about seeking further assistance.

In conclusion, their RAG -augmented LLM chatbot model, powered by Llama-2, holds significant potential for providing support in sensitive contexts such as sexual harassment assistance. Future directions for enhancing this model include implementing web scraping to maintain an updated resource database, transitioning to a cloud-based infrastructure to increase accessibility, and incorporating multi-language support to serve a broader population. These advancements could increase the chatbot's efficacy and reach, allowing it to better support victims of sexual harassment in India and potentially beyond.

3.3 Yang et al. (2024)

Yang et al. (2024) propose ChatDiet, a novel framework for nutrition-oriented food recommendation chatbots designed to deliver highly personalized and interactive dietary guidance. The proposed architecture demonstrates an effective blend of general language understanding with user-specific needs, showcasing the capabilities of LLMs in specialized recommender tasks.

ChatDiet's design addresses three major limitations in existing food recommender systems: personalization, explainability, and interactivity. Traditional nutrition recommender systems often fail to account for individual physiological needs and rely heavily on population-level standards, limiting their ability to provide truly personalized recommendations. ChatDiet's approach utilizes both a "Personal Model" and a "Population Model" in conjunction with ChatGPT 3.5 to overcome these constraints. The Personal Model incorporates individual-specific data, such as personal food preferences, health metrics, and causal effects of nutrition on health, gathered from wearables and digital health records. This model constructs a personalized dietary profile, including causal relationships that dictate how different nutrients affect the user's unique health outcomes. In contrast, the Population Model supplies generalized nutritional data derived from food databases, which provides a foundational knowledge base for the LLM, which generates the response with all the previous information.

The main point of the proposed architecture is the orchestrator module responsible for retrieving, filtering, and transcribing relevant data from the personal and population models, illustrated by Figure 8. This orchestrator preprocesses user queries to ensure the ChatGPT 3.5 receives only relevant inputs, enhancing response accuracy. The retrieval process is equipped with the Best Match 25 (BM25) retrieval algorithm (ROBERTSON; ZARAGOZA, 2009), which is based on the tf-idf ranking function and combines elements of term frequency-inverse document frequency (TF-IDF) and probabilistic models.

The orchestrator also performs prompt engineering to guide the LLM in formulating recommendations in a step-by-step manner, using Zero-Shot Chain-of-Thought prompt engineering methods (WEI et al., 2022), a feature that strengthens explainability by mak-

ing each recommendation's basis clear to users. Through this structure, the framework achieves high interactivity, allowing it to respond to user feedback in real-time, incorporate new dietary preferences, and adjust recommendations dynamically based on evolving user needs.

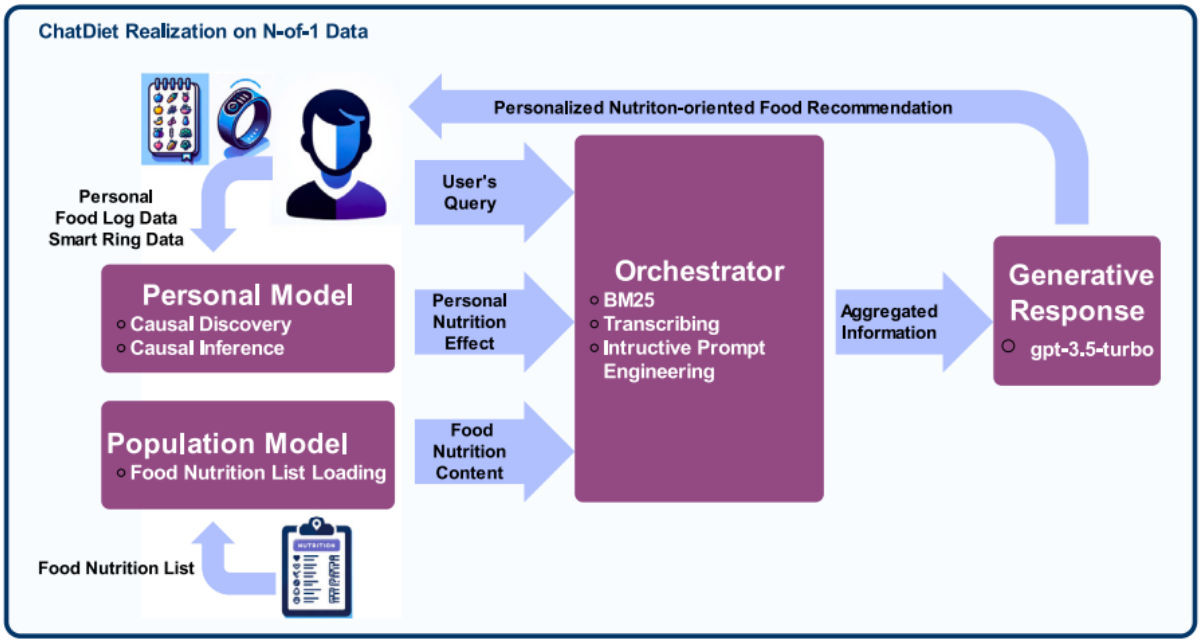

Figure 8 – Overview of ChatDiet's architecture (YANG et al., 2024).

The authors evaluated ChatDiet using both real-world and synthetic datasets, with an effectiveness test showing a 92% accuracy rate in food recommendations. This outcome emphasizes the chatbot potential in aligning dietary suggestions with individual health needs, underlining the importance of causal inference in personalizing health interventions. Dialogue examples, like Figure 9, demonstrate a personalized response of an user query asking for a food recommendation, included in the study. This example further illustrates ChatDiet's ability to generate explainable and logically coherent recommendations, improving user trust and engagement.

In conclusion, Yang et al. (2024) contribute a robust framework for a personalized, LLM-powered chatbot in the nutrition domain, demonstrating the broader applicability of ChatGPT 3.5 in health-focused recommender systems. The study's architecture, combining personalized and population-level models with an LLM, provides insights that can extend to other health-oriented chatbot applications, including mental health, where similar requirements for personalization, sensitivity, and contextual relevance are crucial.

3.4 Qiu et al. (2024)

On another perspective, Qiu et al. (2024) introduce SMILE, an innovative method for generating mental health support chat data by transforming single-turn dialogues into

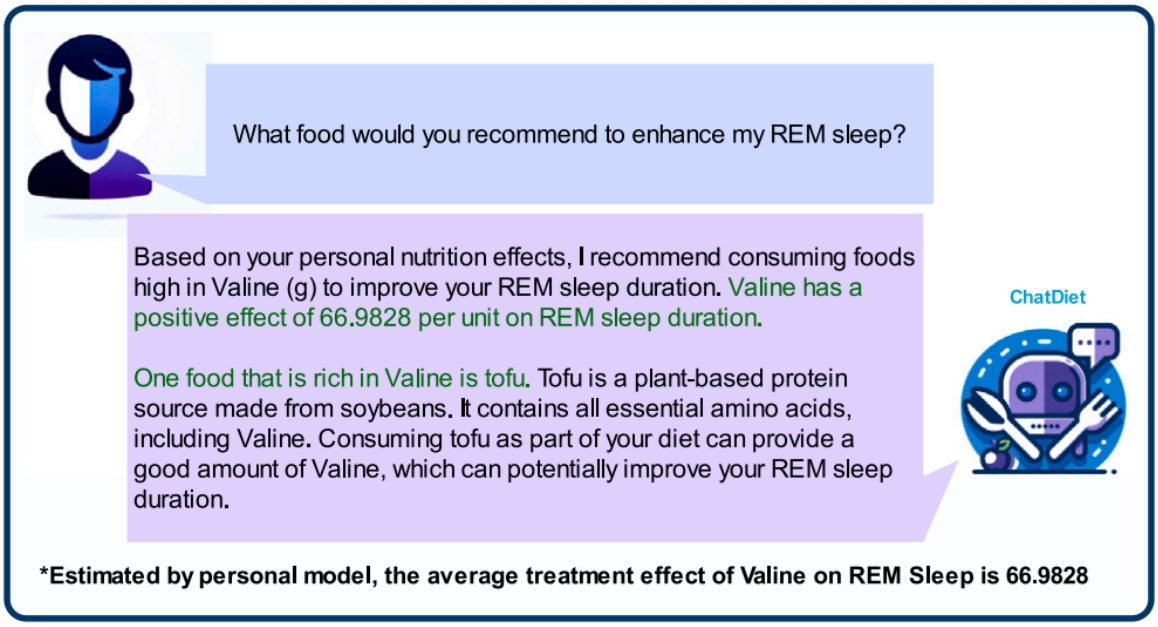

Figure 9 – Example of ChatDiet's personalized recommendation (YANG et al., 2024).

multi-turn conversations. The primary aim of SMILE is to address the scarcity of high-quality, multi-turn conversational datasets essential for developing effective mental health chatbots. This aligns with the broader goal of improving accessibility and personalization in mental health care through scalable and privacy-conscious conversational agents.

The SMILE framework employs ChatGPT to transform single-turn QAs from the Chinese PsyQA dataset into diverse, multi-turn conversations using a role-playing prompt-based engineering approach. To ensure compatibility with the token limitations of ChatGPT, the PsyQA entries are preprocessed through cleaning and truncation while preserving linguistic diversity. The transformation process is guided by carefully crafted prompts that incorporate therapeutic strategies, ensuring that the generated conversations exhibit emotional support, contextual relevance, and professionalism. Hyperparameter tuning, including adjustments to temperature and top-p, further enhances lexical diversity and conversational coherence in the generated dialogues.

The study's architecture integrates data augmentation techniques with a robust evaluation framework to assess the quality of the generated dialogues. SMILE-generated conversations are analyzed using lexical diversity metrics, such as Distinct-n, and semantic measures, like cosine similarity, to validate their authenticity and alignment with real-world dialogue patterns. Human evaluations, conducted using professional feedback and real-life test sets (e.g., PsyTest), reveal that the system effectively maintains empathy, engagement, and linguistic precision. Additionally, fine-tuned dialogue models based on ChatGLM2-6B leverage the SMILECHAT dataset, achieving significant improvements in conversational metrics, such as BLEU and BERTScore, thereby demonstrating the system's utility in mental health applications.

The contributions of the SMILE study include the public release of the SMILECHAT dataset[2], which contains over 55,000 multi-turn conversations, and the development of the MeChat dialogue system, which generates responses using the ChatGLM2-6B language model, fine-tuned with the SMILECHAT dataset, which contains over 55,000 multi-turn dialogues generated with ChatGPT. The Figure 10 illustrates an example of multi-turn dialogue generated by SMILE method in a English translated version.

These advancements not only address the data scarcity in mental health applications but also provide a scalable framework for generating multi-turn conversational datasets in other domains, such as legal and financial advice. By focusing on fine-tuning and prompt engineering, the findings emphasize the potential of LLMs to create scalable, empathetic, and contextually adaptive conversational agents. This work reinforces the utility of LLMs in sensitive domains like mental health support, offering a path toward more inclusive and effective dialogue systems.

Help-seeker: I'm feeling lost after finishing high school, and I'm scared about the college entrance exam. Can you give me some advice?
Supporter: You've identified your concerns. Don't worry; we'll work through this together. You're likely under immense pressure during the college entrance exam period. When was the last time you took a moment to be with yourself and relax?
Help-seeker: It's been a long time since I took care of myself properly. At this stage, parents, teachers, and classmates around me are all dealing with their anxieties. I feel really sad, like nothing matters.
Supporter: I understand how you feel. At this age, we sometimes feel denied, ignored, or even redundant. But please believe that you have value to the world and others.
Help-seeker: Yes, I need to rediscover my sense of self-worth. I'm living a very tiring life now, almost no friends, always alone. I don't know how to face the college entrance exam.
Supporter: The college entrance exam is indeed a significant challenge, but we shouldn't let it completely define our value. We need to learn to adjust our emotions, not be easily influenced by others, and not blindly prove ourselves. Let's calm down and think about how to improve ourselves.
Help-seeker: What you're saying is right; the college entrance exam is just a phase. But I feel lost; I don't know how to study, and every exam is a setback for me.
Supporter: Learning is a specific skill, and we can learn together how to manage time, rest, and even what to do during breaks. Moreover, we can continuously improve our study methods through communication with classmates and teachers. As long as we maintain a positive mindset, progress is possible.
Help-seeker: Thank you for your encouragement and advice. I feel like I need more support and understanding, not to feel lonely and helpless anymore. Do you have any other suggestions?
Supporter: When you feel sad and lost, look for positive things to motivate yourself, like reading inspirational books or watching uplifting videos. Also, don't forget to communicate with friends; their support and encouragement will help you get through tough times.
Help-seeker: Yeah, I'll try to find some positive things to give myself motivation. I'm really grateful for your companionship and help.
Supporter: It's okay; I'll be with you throughout the college entrance exams. Believe in yourself; you can overcome difficulties and achieve good results. Keep going!
Help-seeker: Thank you for your trust and encouragement. I'll do my best!

Figure 10 – Example of a multi-turn dialog generated by SMILE method (QIU et al., 2024).

3.5 Sanna et al. (2024)

Sanna et al. (2024) addresses the challenges and innovations involved in developing certified medical chatbots capable of delivering accurate, consistent, and certified information. The proposed system combines a modular RAG framework with LLMs to overcome

[2] Available at: <https://github.com/qiuhuachuan/smile>

limitations associated with unstructured datasets and ensure conversational adaptability. This approach is particularly relevant in contexts where the precision and reliability of responses are paramount, such as healthcare.

The methodology centers around a modular RAG architecture that integrates certified knowledge with advanced conversational capabilities. A key component of the system is the use of Hypothetical Document Embeddings (HyDE), which enhances the retrieval process by generating intermediary hypothetical documents that align user queries with relevant certified documents. These hypothetical documents, generated by LLMs , serve as a bridge between the user's intent and the certified content. Despite potential hallucinations in the hypothetical documents, the method ensures that the certified answers retrieved from the knowledge base remain accurate and verifiable.

The workflow begins with the preparation of a certified dataset sourced from trusted medical repositories. Texts are segmented into manageable chunks to facilitate conversational integration. This segmentation process ensures that the responses remain concise and contextually appropriate, adhering to the requirements of conversational systems. For each user query, the system first generates a hypothetical document using GPT-4-turbo. This document is then used to retrieve relevant certified information from a knowledge base through a combination of Bi-Encoder and Cross-Encoder models. While the Bi-Encoder quickly identifies potential matches based on semantic similarity, the Cross-Encoder refines these matches by re-ranking the results to ensure the most relevant documents are selected.

To maintain the conversational quality of the chatbot, the retrieved documents are summarized into brief responses of 80–120 words. These summaries are further augmented with guardrail mechanisms to ensure compliance with medical certification standards and brevity requirements. The final response provided to the user combines this concise summary with references to the original certified documents, enhancing transparency and trust. The Figure 11 illustrates the full pipeline.

The system was evaluated using a set of 100 user-generated queries, achieving an 85% success rate in retrieving relevant documents. This performance is a significant improvement over traditional intent classification models, such as RASA[3], which managed only a 13% accuracy rate. The HyDE-based approach demonstrated its robustness in addressing diverse and complex queries while minimizing the risk of inaccuracies. Despite these successes, some challenges remain, including difficulties associated with the stylistic and semantic heterogeneity of unstructured data, which can occasionally hinder effective topic modeling.

In conclusion, the study highlights the effectiveness of combining modular RAG frameworks with advanced LLMs in building reliable and empathetic chatbots for medical applications. By addressing key challenges such as hallucinations and data heterogeneity,

[3] Available at: <https://github.com/RasaHQ/rasa>

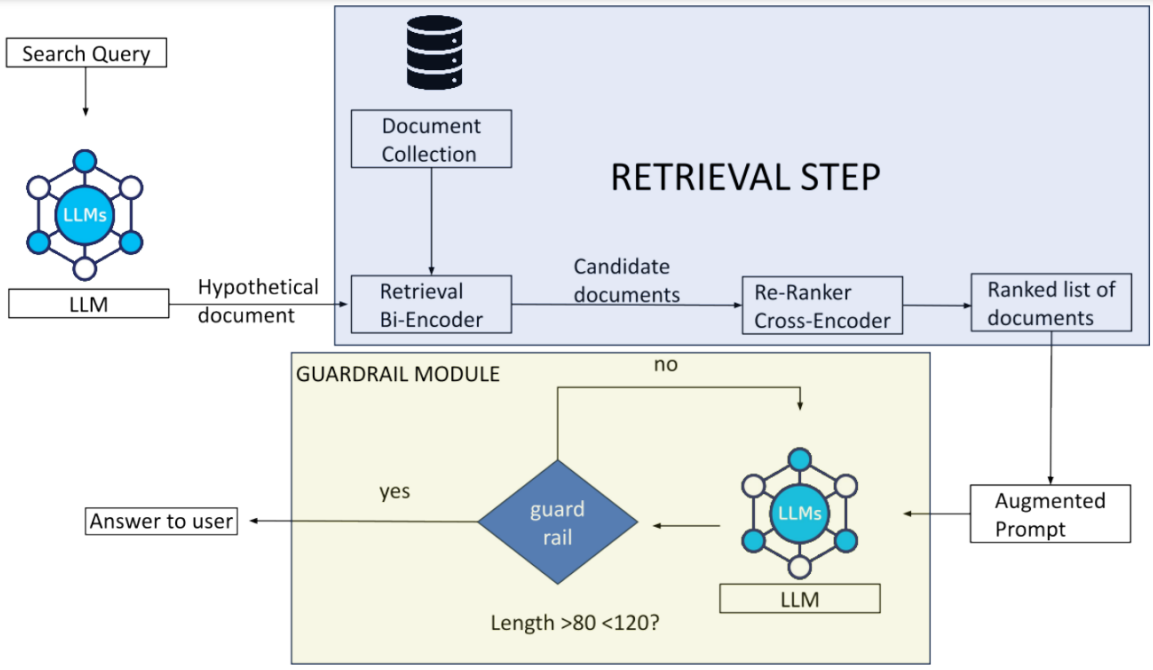

Figure 11 – Overview of the RAG framework by Sanna et al. (2024).

the proposed methodology sets a strong foundation for the development of certified conversational systems. However, the authors acknowledge areas for further improvement, including expanding multilingual capabilities and refining data annotation processes to enhance adaptability and scalability in real-world applications.

3.6 Fan et al. (2024)

In parallel, introduced a knowledge-guided mental health chatbot framework designed to provide personalized psychological counseling. The chatbot leverages pre-trained LLMs, such as Qwen2-7b, GPT-4, and GPT-3.5, alongside domain-specific knowledge bases, using fine-tunning (FT) and RAG technics offering tailored interventions for various mental health conditions. The framework has two main modules that operate based on RAG: a mental disorder classification model based on psychological counseling, dialogues and a conversational robot model, illustrated in Figure 12.

The classification module employs the Qwen2-7B model fine-tuned with psychological counseling dialogues, from a classification dataset (LIU et al., 2021b) and a conversational one (LIU et al., 2021a), to identify mental health conditions such as depression, anxiety, PTSD, and bipolar disorder. To improve diagnostic precision, the system retrieves structured knowledge from a DSM-5-aligned knowledge graph, where nodes represent symptoms and disorders, and edges define semantic relationships. The retrieval process utilizes a similarity-based matching function to extract the most relevant information from the graph, which is then integrated into the Qwen2-7B (FT+RAG) model for classification, ensuring context-aware assessments.

After classification, the chatbot employs a hierarchical text retrieval method to generate treatment recommendations, leveraging a structured psychological treatment manual. The retrieval module iteratively searches across different levels of treatment guidelines, from general interventions to more specialized therapeutic strategies. The retrieval-enhanced fine-tuning (FT+RA) approach allows the chatbot to dynamically refine responses by continuously re-evaluating classification outputs and adjusting treatment recommendations based on real-time interactions. Additionally, knowledge integration is performed through a large-scale prompt engineering framework, ensuring that the chatbot aligns its recommendations with established psychotherapeutic methodologies.

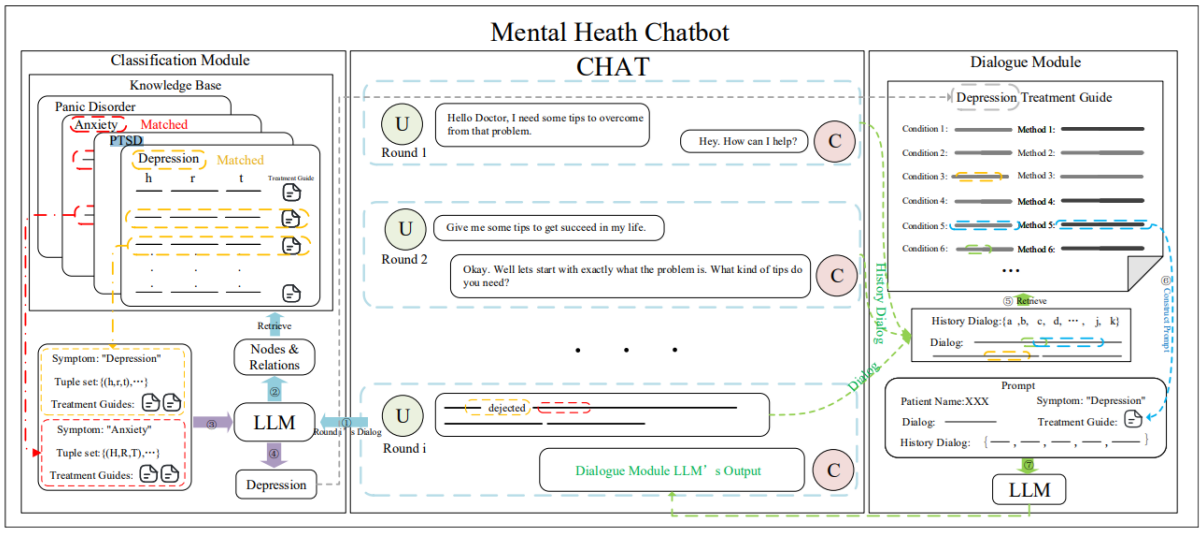

Figure 12 – Overview of the RAG framework by Fan et al. (2024).

The framework's evaluation employed diverse metrics to assess its efficacy across classification and conversational tasks, using Word2Vec, SBERT, and RoBERTa similarity metrics, as well as GPT-4-based scoring mechanisms to assess fluency, empathy, and expertise, on diverse combinations of Qwen2-7b, ChatGPT-3.5, ChatGPT-4.0 with fine-tunning and RAG. For the classification module, the fine-tuned Qwen2-7b model with retrieval augmentation achieved a notable accuracy of 74.70%, significantly outperforming its baseline counterpart, which recorded an accuracy of 50.60%. Precision and recall metrics also reflected substantial improvements, with the fine-tuned model reaching 78.15% precision and 69.64% recall. For conversational quality, human evaluators assigned average scores of 72.50 for professionalism, fluency, and empathy when using the retrieval-augmented model, slightly surpassing GPT-4's score of 71.33. Additionally, cosine similarity metrics confirmed superior alignment between generated responses and reference texts, with values reaching 0.8826 using SBERT.

Further experiments highlighted the efficiency gains facilitated by retrieval augmentation. Large models such as GPT-4 exhibited reduced inference times when integrated with retrieval mechanisms, averaging 2.12 seconds per query compared to 3.45 seconds without

retrieval augmentation. This improvement underscores the scalability and practicality of the proposed system, particularly in resource-constrained environments.

3.7 Li et al. (2024)

Li et al. (2024) explores the development of LD-Agent, a model-agnostic framework designed for long-term open-domain dialogues, addressing challenges in maintaining event memory, persona consistency, and cross-domain adaptability. Unlike traditional dialogue systems that focus on single-session interactions, LD-Agent supports extended multi-session dialogues by dynamically integrating historical events and personas to ensure coherent and personalized responses over time.

The methodology is centered on three key modules: the event memory module, the persona extraction module, and the response generation module. The event memory module distinguishes between long-term and short-term memory. Long-term memory stores summarized historical events as vector representations, enabling efficient retrieval for coherence across sessions. These summaries are enhanced through instruction-tuned event summarization, improving the quality of stored memories. Short-term memory maintains contextual details of the ongoing session, which are periodically summarized and transferred to long-term memory to ensure scalability. The framework is shown by figure 13.

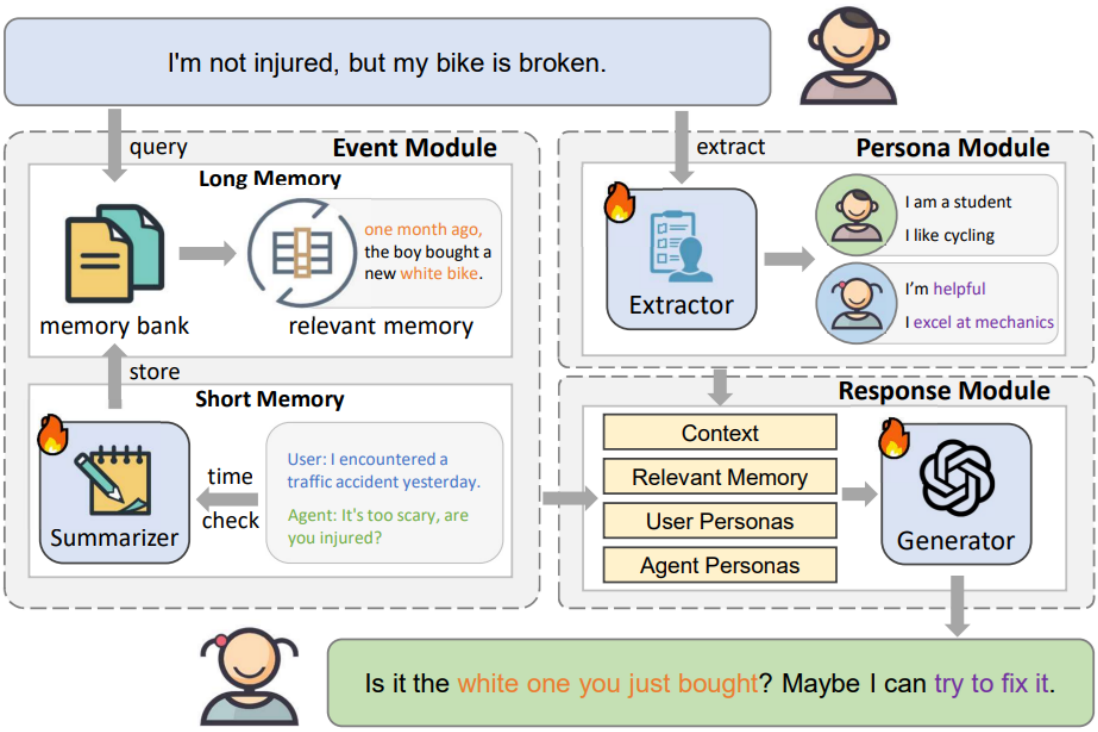

Figure 13 – Overview of the memory-based framework by Li et al. (2024).

The persona extraction module dynamically models user and agent personas using a

bidirectional approach, using Zero-Shot and Chain-of-Thought prompting. By employing LoRA-based instruction tuning, the system extracts personality traits from dialogue utterances and stores them in separate persona banks. This ensures consistency in the chatbot's responses while adapting to evolving user interactions. The response generation module synthesizes retrieved memories, persona data, and current session context to produce contextually appropriate and personalized responses.

Evaluation was conducted on two benchmark datasets — MSC and Conversation Chronicles (CC) — featuring long-term multi-session dialogues. The LD-Agent framework was tested on several baseline models, including ChatGPT, ChatGLM, and BlenderBot, in both zero-shot and fine-tuned settings. Automatic evaluation metrics such as BLEU-N, ROUGE-L, and METEOR assessed response fluency and coherence, while human evaluations measured topic coherence, fluency, and engagingness. Results showed that LD-Agent consistently outperformed existing models, achieving significant improvements across all metrics. For example, in BLEU-2 scores, LD-Agent-enhanced models exhibited gains of up to 10% over their baseline counterparts, with robust performance across multiple sessions.

The study highlighted the critical role of the event memory module, which contributed the most to maintaining dialogue coherence across sessions. Persona extraction modules also improved the personalization and consistency of responses, particularly in scenarios involving extended user-agent interactions. Additionally, cross-domain evaluations demonstrated LD-Agent's adaptability, with models tuned on one dataset performing competitively on another, affirming the framework's generality.

In conclusion, LD-Agent provides a scalable and adaptable solution for long-term dialogue systems, leveraging modular components to enhance coherence, personalization, and adaptability. Its superior performance on multi-session tasks underscores its potential for real-world applications, though future work could address limitations related to dataset size and the complexity of longer dialogue scenarios, representing a advance in personalized, long-term conversational agents.

3.8 Abbasian et al. (2024)

In a study focusing on personalization, Abbasian et al. (2024) introduce "openCHA", a novel framework powered by LLMs designed to create CA with advanced problem-solving and personalized capabilities. The authors addresses the limitations of current CAs, such as their inability to handle complex healthcare tasks, integrate multimodal data, and deliver highly personalized interactions. The "openCHA" combines LLM capabilities with external data sources, knowledge bases, and advanced AI models to generate accurate, contextually relevant, and trustworthy responses to user queries.

The framework employs state-of-the-art models, including ChatGPT for conversa-

tional capabilities, BioGPT for biomedical text generation and analysis, and domain-specific LLMs like Med-PaLM for retrieving and synthesizing healthcare knowledge. These models are integrated to leverage their specialized expertise in generating reliable responses based on medical information.

A key feature of openCHA is its robust retrieval mechanism, designed to ensure that responses are grounded in up-to-date and accurate information. The system accesses external knowledge bases, such as reputable medical literature, knowledge graphs, and curated healthcare datasets. Using advanced retrieval models and search APIs, the framework extracts relevant information to address user queries. For instance, Google Search APIs and specialized retrieval tools are employed to identify the most pertinent data, which is then validated and synthesized into user-friendly responses. This retrieval process minimizes hallucination risks by relying on certified data sources and integrating them with real-time analysis tools, ensuring that responses remain both accurate and reliable.

Personalization is another central focus of the framework. openCHA integrates user-specific data, including longitudinal health records, biosignals, and demographic information, to tailor responses. For instance, the system can analyze a user's heart rate variability data to estimate stress levels or assess physical activity metrics to generate detailed health reports. This integration ensures that responses are not only accurate but also relevant to the user's unique health profile. All the framework is managed by an orchestrator acting as a problem solver to address healthcare-related queries by analyzing input queries, gathering the required information, performing actions, and offering personalized responses, as it shown the Figure 14. All the retrieved information is incorporated using the Tree of Thought (YAO et al., 2023) prompting method, which 1) generates three unique strategies (i.e., sequences of tasks to be called with their inputs), 2) describes the pros and cons of each strategy, and 3) selects one as the best strategy.

Evaluation metrics highlight the framework's effectiveness in healthcare applications. Through two demonstrations and multiple use cases, openCHA showcases its ability to handle complex healthcare tasks. In one demo, the framework analyzes patient sleep and activity data, achieving precise summaries and follow-up recommendations. In another, it estimates stress levels using biosignal data with 86% accuracy, demonstrating its capability to process multimodal inputs. Additionally, use cases like ChatDiet (YANG et al., 2024) and diabetic patient management illustrate openCHA's performance in providing tailored recommendations and health management insights.

In summary, openCHA provides a comprehensive framework for developing CA capable of integrating multimodal data, leveraging external sources, and ensuring personalized interactions. While promising, the framework faces challenges related to response latency, token limits, and privacy concerns, particularly in handling sensitive healthcare data. Addressing these issues will be critical to enhancing its scalability and applicability in real-world healthcare scenarios.

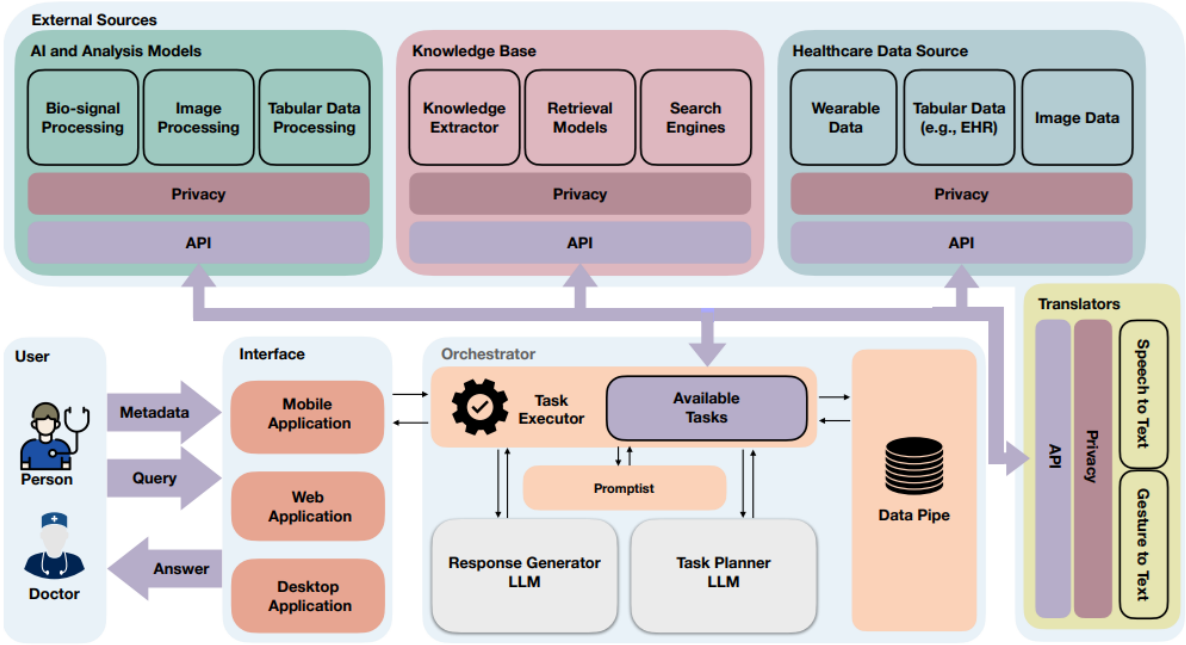

Figure 14 – Overview of the Orchestrator-based Chatbot by Abbasian et al. (2024).

3.9 Jafari et al. (2025)

In a arabic study, Jafari et al. (2025) propose a psychological health chatbot designed to facilitate conversational support for Persian-speaking individuals experiencing mental health challenges. The chatbot aims to provide emotionally aware and contextually relevant interactions, assisting users in expressing their concerns while ensuring safe and appropriate engagement. By tailoring the system to a low-resource language, this study expands AI-driven mental health interventions beyond dominant linguistic markets.

The chatbot integrates multiple NLP components, including emotion detection and stress classification, to personalize user interactions. While the classification module relies on ParsBERT and XLM-RoBERTa, the core conversational engine is powered by ChatGPT-4.0 Mini. The response generation process follows a structured pipeline described in Figure 15: user messages are first analyzed for emotion and psychological distress, then contextualized within the conversation history, and finally used to construct a customized prompt for the LLM. The prompt includes details on the user's detected emotional state, conversational history, and an instruction set directing the model to generate empathetic and psychologically appropriate responses.

To enhance response coherence, the chatbot assigns weighted relevance scores to previous messages, ensuring that responses remain aligned with ongoing discussions. Additionally, the system employs a validation mechanism that filters outputs for toxicity and emotional appropriateness, leveraging XLM-RoBERTa-large for automated content moderation. This safeguards against responses that could inadvertently escalate distress or provide misleading advice, reinforcing the chatbot's role as a supportive rather than

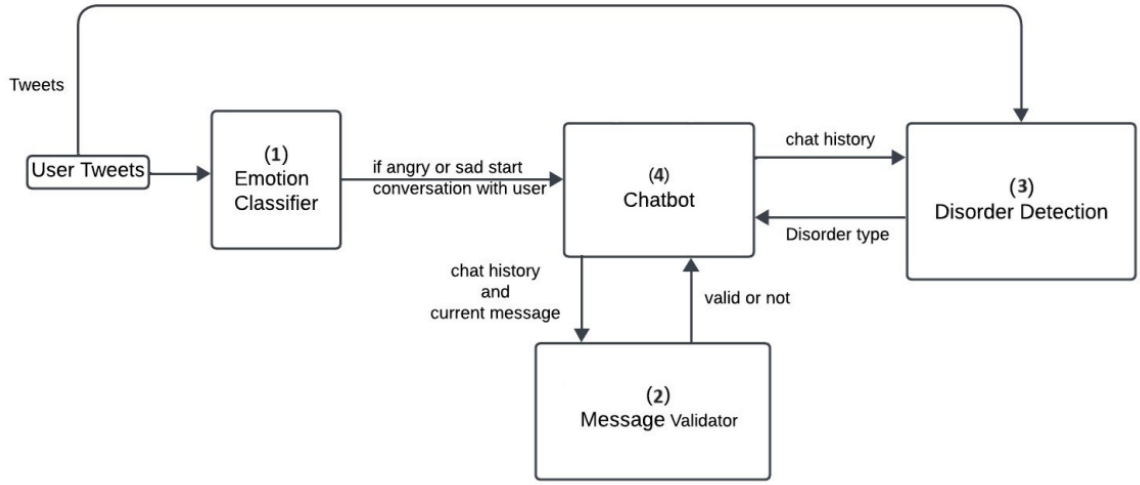

Figure 15 – The structure of the mental health conversational agent by (ABBASIAN et al., 2024).

diagnostic tool.

Evaluation of the chatbot involved qualitative assessments. The PHQ-9 depression scale was used to measure changes in participants' emotional well-being before and after chatbot interactions over a two-week period, revealing that 7 out of 9 users experienced emotional improvement. Furthermore, 7 out of the 9 participants exhibited signs of improvement by the end of the experiment.

Despite these advancements, the study acknowledges areas for improvement, particularly in handling complex emotional states and refining long-term conversational memory. Future work aims to integrate Retrieval-Augmented Generation (RAG) for improved response contextualization and expand the chatbot's multilingual capabilities to serve a broader population. These efforts underscore the potential of conversational AI in bridging mental health accessibility gaps while ensuring ethical and effective interactions.

3.10 Zhao et al. (2024)

Unlike traditional approaches focused on building Emotional Support Conversation (ESC) systems, Zhao et al. (2024) introduced a novel evaluation framework, ESC-Eval, aimed at systematically assessing the performance of LLMs in providing emotional support. This framework represents a shift from the development of ESC chatbots to the rigorous evaluation of their capabilities. By leveraging role-playing agents to simulate human distress scenarios, ESC-Eval provides a robust and scalable methodology that addresses the limitations of conventional human and statistical evaluation methods.

ESC-Eval employs a novel combination of pre-constructed role-playing cards and a dedicated role-playing model called ESC-Role. The role cards, extracted and refined from seven datasets, encompass diverse real-world scenarios such as mental health struggles,

academic pressures, and interpersonal issues. ESC-Role, fine-tuned on 14,000 instances of role-playing dialogues, demonstrates enhanced capabilities in mimicking realistic human behavior compared to state-of-the-art models like GPT-4 and BaichuanNPC. These components enable the systematic evaluation of ESC models across seven dimensions, including empathy, fluency, thematic consistency, and human-like attributes.

The study evaluated 14 LLMs, including general-purpose AI assistants (e.g., GPT-4) and domain-specific ESC models (e.g., ExTES-Llama). Key metrics such as fluency, diversity, and empathy were assessed through both manual annotation and the automated scoring system ESC-RANK. Notably, the domain-specific EmoLLM achieved the highest overall performance in Chinese dialogues, with an average score of 71.69, while general-purpose models excelled in fluency and suggestion effectiveness. Pairwise comparisons revealed that ESC-Role's dialogues were nearly indistinguishable from human-generated interactions, achieving a win rate comparable to human performance in terms of conversational realism.

ESC-Eval also introduced the automated scoring model ESC-RANK, trained on over 59,000 annotated dialogues. This model surpassed GPT-4 by 35 points in scoring accuracy, achieving over 99% accuracy within a one-point deviation range. ESC-RANK's integration ensures an efficient, scalable evaluation process for future ESC models, reducing reliance on extensive human annotations while maintaining robust assessment quality.

In summary, Zhao et al. (2024) provide a new framework for evaluating ESC models, setting a new benchmark for the field. While the study highlights the superior performance of domain-specific models, it also underscores the need for further advancements in emotional support knowledge and human-centric responses. Future research should focus on enriching datasets, refining role-playing models, and enhancing the interpretability and personalization of ESC systems.

3.11 Ye et al. (2025)

Ye et al. (2025) present SweetieChat, a strategy-enhanced role-playing framework designed to improve emotional support agents in open-domain settings. The study starts from two recurrent limitations observed in general-purpose LLMs: responses often become verbose and formulaic, and the model tends to overuse a small subset of comforting strategies, reducing conversational depth and limiting its ability to adapt across heterogeneous user situations. To address these issues, the authors propose constructing a larger and more diverse emotional support corpus through role-playing, while explicitly controlling which emotional support strategy should guide each assistant turn.

The proposed framework separates the dialogue generation process into three explicit roles with different responsibilities, so that diversity and strategic behavior are induced by construction rather than expected to emerge implicitly from a single model. The roles

are:

- ❏ Seeker: simulates the user who is seeking support. The Seeker is defined by a problem type, a concrete scenario description, and a character profile. To broaden coverage, the framework expands the set of seeking problem types beyond those originally defined in ESConv, then uses a seed pool of high-quality examples to generate new scenarios that describe specific events in sufficient detail. A second seed pool is used to generate seeker profiles, allowing the same type of situation to be expressed by individuals with different backgrounds.

- ❏ Strategy Counselor: selects the support strategy that should be used in the next assistant response. The Counselor is therefore trained as a strategy selector: given the dialogue history up to the current user turn, it outputs one strategy label from the predefined set used in ESConv. Concretely, the Counselor is obtained by instruction fine-tuning a LLaMA model on ESConv to map a dialogue prefix to the most appropriate strategy, aiming to enforce more deliberate transitions such as questioning during exploration and suggestions after sufficient understanding has been established.

- ❏ Supporter: generates the assistant response conditioned on the chosen strategy. This role is responsible for producing the actual emotional support utterance, but it is constrained by the strategy proposed by the Counselor. The Supporter is instantiated with a strong LLM and prompted with the dialogue history, the Seeker configuration, and the selected strategy, so that the generated response follows the intended counseling move rather than defaulting to generic reassurance.

Through iterative interactions between Seekers from diverse scenarios and the Supporter, guided turn-by-turn by the Strategy Counselor, the authors construct the ServeForEmo dataset. After post-processing and quality filtering, ServeForEmo contains 3,757 multi-turn dialogues and 62,863 utterances, and is used as supervised data to fine-tune a LLaMA-based emotional support agent (SweetieChat). Evaluation is performed on the manually annotated ESConv test set using automatic metrics (BLEU, ROUGE-L, Distinct-n, and BERTScore) as well as human preference judgments. The reported results indicate that fine-tuning on ServeForEmo improves semantic alignment and response diversity relative to baselines trained on earlier ESC data augmentation methods and on ESConv-only supervision.

Therefore, this work supports the premise that emotional support generation benefits from explicit control of conversational strategies, reinforcing the importance of structured prompting and controlled response planning in ESC settings. Secondly, SweetieChat demonstrates that user variation can be operationalized through explicit seeker profiles

and scenario diversification; this motivates the investigation of persona conditioning in the current study.

3.12 Kermani, Perez-Rosas and Metsis (2025)

Kermani, Perez-Rosas and Metsis (2025) present a systematic comparison of three deployment strategies for large language models in mental health text analysis: prompt engineering, retrieval-augmented generation (RAG), and fine-tuning. The central purpose of the study is to clarify the practical trade-offs among these strategies when applied to clinically sensitive text classification tasks, where reliability is critical and computational constraints often limit the feasibility of full adaptation. To ensure a controlled comparison, the same backbone model is used across all approaches, namely the 8B parameter version of LLaMA 3, and performance is reported on two complementary classification problems: emotion classification and mental health condition detection.

The experimental design relies on two public datasets that capture different dimensions of mental health-related language. The first dataset is the DAIR-AI Emotion dataset, containing 20,000 tweets annotated with six basic emotions (joy, sadness, anger, fear, love, surprise) and split into train/validation/test partitions. The second dataset is the Reddit SuicideWatch and Mental Health Collection (SWMH), containing 54,412 Reddit posts labeled with four mental health condition categories (depression, anxiety, bipolar disorder, suicidal ideation). A preprocessing pipeline is applied to standardize the inputs, including removal of URLs and mentions, UTF-8 normalization, truncation of long texts to the model token limit, and label-quality checks.

For fine-tuning, the work adopts parameter-efficient adaptation via LoRA and applies 4-bit quantization to reduce memory usage, targeting a configuration that remains feasible on constrained hardware. In the prompt-engineering approach, two variants are considered: zero-shot setting uses a direct instruction that lists the available labels and requires the model to output exactly one label, and few-shot setting, which extends the same prompt with labeled examples selected from the training split, using two examples per label to balance class coverage. In the RAG strategy, retrieval is used to provide in-context examples at inference time rather than to inject external domain guidelines. At inference time, the input is embedded, top-k nearest neighbors are retrieved with an explicit attempt to include diverse examples across categories, and the three most similar retrieved examples are assembled into a unified context block appended to the classification prompt.

Empirically, fine-tuning yields the strongest performance on both tasks, reaching 91% accuracy (macro-F1 0.87) on DAIR-AI Emotion and 80% accuracy (macro-F1 0.81) on SWMH. Among non-fine-tuned alternatives, zero-shot prompting is reported as the strongest, achieving 49% accuracy (macro-F1 0.38) on DAIR-AI Emotion and 68% ac-

curacy (macro-F1 0.67) on SWMH, surpassing both few-shot prompting and RAG in this setting. Few-shot prompting achieves 39% accuracy (macro-F1 0.30) on DAIR-AI Emotion and 45% accuracy (macro-F1 0.57) on SWMH, while RAG achieves 47% accuracy (macro-F1 0.32) and 56% accuracy (macro-F1 0.45), respectively. The analysis emphasizes that RAG outcomes are highly dependent on retrieval quality, with substantial degradation when retrieved examples are weakly related to the input, reinforcing the need for careful retriever design and validation when RAG is deployed for mental health tasks.

While Kermani, Perez-Rosas and Metsis (2025) focuses on classification rather than response generation, the findings support two design choices that are central to the current study: (i) fine-tuning is a strong option when computational resources permit domain adaptation, and (ii) retrieval pipelines require robust retrieval mechanisms and curated evidence sources to avoid performance collapse under noisy or weakly relevant context, which motivates hybrid retrieval and guideline-based knowledge bases in generation-oriented settings.

3.13 Zerhoudi and Granitzer (2026a)

Zerhoudi and Granitzer (2026a) introduce PersonaRAG, a personalized retrieval-augmented generation framework that augments standard RAG with user-centric agents to adapt retrieval and answer formulation to the user's contextual needs. The work is motivated by two limitations of conventional RAG: retrieved passages may contain noise or be weakly aligned with the user's intent, and the retrieval process typically ignores user context, producing answers that can be correct but not tailored to what the user actually needs. PersonaRAG addresses this gap by leveraging live interaction signals and user-profile information to continuously refine retrieval decisions and adjust the final response in a transparent manner.

The methodology is organized as a three-step pipeline: retrieval, user interaction analysis, and cognitive dynamic adaptation. In the retrieval step, the system retrieves top-k passages for a user query using a standard sparse retriever (BM25) over an external corpus. In the user interaction analysis step, the system decomposes personalization into multiple specialized agents that interpret different aspects of user behavior and preferences. The User Profile Agent manages and updates profile signals from historical interactions; the Contextual Retrieval Agent proposes refinements to the query and retrieval priorities using the profile; the Live Session Agent models short-term session intent from real-time behaviors; the Document Ranking Agent re-ranks retrieved passages by combining long-term and session-level signals; and the Feedback Agent aggregates implicit and explicit feedback to improve subsequent ranking decisions. These intermediate outputs are consolidated into a shared global memory (Global Message Pool) that supports consistent

inter-agent communication. In the cognitive dynamic adaptation step, an initial answer is produced (including a chain-of-thought baseline in their comparisons), and a Cognitive Agent revises the answer using the agent-generated insights to better align content, detail level, and emphasis with the user's inferred information needs.

Evaluation is conducted on single-hop question answering tasks using NaturalQuestions, TriviaQA, and WebQuestions, with 500 randomly sampled queries per dataset to control API costs. The experiments compare PersonaRAG against non-RAG baselines and several RAG variants (including direct passage injection and self-refinement approaches). Effectiveness is measured primarily via accuracy using exact string matching between predictions and gold answers, while user-centric adaptability is analyzed using BLEU-2 similarity between outputs from different configurations, complemented by post-hoc analyses of average sentence length and syllable count. Across datasets, PersonaRAG improves over conventional RAG settings, with particularly large gains on WebQuestions when using Top-3 and Top-5 passages. The paper also reports that PersonaRAG's principles can be transferred to other LLM backbones by using PersonaRAG outputs to guide open-source generators (e.g., LLaMA3-70B and MoE-8x7b), which yields consistent accuracy improvements in their reported setup. The authors note practical trade-offs, including increased latency and cost due to multiple LLM calls and longer prompts that include both retrieved passages and user-centric analyses.

PersonaRAG supports the design choice of conditioning both retrieval and generation on user profile information, which is conceptually consistent with the persona oriented model variant implemented in this thesis. However, the target task differs: PersonaRAG focuses on open-domain question answering with live interaction signals, whereas the present study applies persona-conditioned retrieval and generation to emotional support dialogue grounded in clinical guideline documents, using predefined personas and structured metadata instead of continuous interaction logs.

3.14 Summary of presented works

The main features of the previous mentioned related works are summarized in Table 1.

Table 1 – Overview of the Selected Studies on LLM-based Conversational Chatbots

Article	**Year**	**Language**	**Models Evaluated**	**Strategy**	**Method of Evaluation**
(YU; MCGUINNESS, 2024)	2024	English	DialogGPT + ChatGPT 3.5	Fine-tunning	Perplexity, BLEU Score and Human Evaluation
(VAKAYIL et al., 2024)	2024	English	Llama-2	RAG	Accuracy
(YANG et al., 2024)	2024	English	ChatGPT 3.5	RAG	Accuracy
(QIU et al., 2024)	2023	Chinese	ChatGPT 3.5	Fine-Tunning	BLEU, METEOR, ROUGE, BERTScore, Distinct-n
(SANNA et al., 2024)	2024	Italian	ChatGPT 4-PRO	RAG	Accuracy
(ZHAO et al., 2024)	2024	English and Chinese	Diverse LLms	-	ESC-Eval
(FAN et al., 2024)	2024	Chinese and English	GPT 3.5 and Qwen2	RAG	GPT-4 Score
(LI et al., 2024)	2024	English	GPT 3.5, ChatGLM3, BART	Retrieval	BLEU-N, ROUGE-L, METEOR, Accuracy
(ABBASIAN et al., 2024)	2024	English	GPT 3.5	RAG	Accuracy
(JAFARI et al., 2025)	2025	Arabic	GPT 4.0-mini	Fine-tunning	Qualitative PHQ-9 Questionary
(YE et al., 2025)	2025	English	LLaMA + GPT-4o	Fine-tunning	BLEU, ROUGE-L, Distinct-n, BERTScore, Human Evaluation
(KERMANI; PEREZ-ROSAS; METSIS, 2025)	2025	English	LLaMA 3	Fine-tunning / RAG	Accuracy, F1-score, Precision, Recall
(ZERHOUDI; GRANITZER, 2026a)	2026	English	GPT-3.5, LLaMA3, MoE-8x7b	Personalized RAG (multi-agent)	Accuracy, BLEU-2, Avg. sentence length, Avg. syllable count

Chapter 4

Resources and Techniques

This chapter describes the resources and techniques employed in the computational experiments reported in Chapter 5. It presents the ESConv dataset selection and preprocessing procedure (Section 4.1), the prompt engineering strategies adopted to construct the model prompt (Section 4.2), the models selected and the fine-tuning setup (Sections 4.3 and 4.4), the retrieval pipeline and vector database used for RAG execution (Section 4.5), the guardrail mechanism used to handle crisis-related inputs (Section 4.6), and the personas adopted for response personalization (Section 4.7).

4.1 Data Preprocesssing

The ESConv (Emotional Support Conversation) dataset was selected as the primary supervised resource for model adaptation in this thesis because it provides a domain-focused corpus of supportive interactions with adequate size and consistent dialog structure. Originally introduced by Liu et al. (2021a), ESConv comprises 1,053 multi-turn dialogs totaling 31,410 utterances, offering sufficient coverage of seeker–supporter exchanges for fine-tuning and evaluation. Unlike open-domain conversational datasets that primarily emphasize topical breadth and generic coherence, ESConv is specifically designed for emotional support dialog and is annotated with emotional support strategies (e.g. comforting and problem-solving). However, in the present work, only the raw dialog text (utterances) is used during training, and the strategy annotations are not leveraged. ESConv has become a standard benchmark for fine-tuning Large Language Models (LLMs) in the emotional support domain, and it has been adopted in recent state-of-the-art systems, including AugESC (ZHENG et al., 2023) and SweetieChat (YE et al., 2025).

Two complementary preprocessing strategies were adopted in this work to accommo-

date the multi-turn nature of ESConv while mitigating the limited amount of available training data, given the modest size of ESConv relative to typical instruction-tuning corpora. The first strategy converts each dialogue into a set of independent question–answer (QA) pairs by mapping a seeker utterance to the immediately subsequent supporter reply. This QA-style transformation yields a larger number of atomic training instances and simplifies supervision by exposing the model to direct input–output mappings, which can improve optimization stability in low-resource settings. However, because it discards earlier turns, this approach does not explicitly enforce discourse-level consistency or long-range dependencies, and may therefore underrepresent phenomena such as gradual emotional disclosure, clarification, and iterative support.

To address these limitations, a second strategy, termed context-aware data splitting, was employed to preserve conversational context and effectively expand the set of training examples without introducing new content. Concretely, each original multi-turn dialogue is restructured into a sequence of training instances where, at turn t, the model receives the concatenation of all prior rounds (i.e., turns $1, \ldots, t-1$) as context and is trained to generate the supporter response at turn t. This rolling-window formulation increases the number of context-conditioned samples and encourages the model to ground each reply in the evolving dialogue state, which is particularly important for emotional support where appropriate responses depend on previously disclosed feelings, constraints, and attempted coping actions. In addition, by generating multiple prefix-to-next-response examples from the same conversation, context-aware data splitting provides a data-efficient mechanism to leverage scarce multi-turn supervision while maintaining the sequential structure required for coherent, context-sensitive support generation.

Table 2 presents the number of instances in the training, validation, and test splits obtained after applying the two complementary preprocessing strategies.

Table 2 – Dataset split sizes after preprocessing.

Split	**Train**	**Val**	**Test**
Instances	10,191	2,156	2,295

4.2 Instructions and Training Prompt Generation

To improve controllability and behavioral alignment during supervised fine-tuning, a instruction-based prompt template that explicitly defines the model's role, scope, and response constraints was used. This choice is motivated by prior (TOPAL; BOZANTA; BASAR, 2024) evidence that fine-tuned LLM behavior can be highly sensitive to instruction phrasing and prompt structure, affecting both output quality and robustness under prompt variations. In parallel, recent work (CHATTERJEE et al., 2025) on instruction-centric training emphasizes that clearly specified task framing helps reduce unwanted

behaviors (e.g., anthropomorphic claims) and improves adherence to desired response policies when optimizing with limited supervised data. Accordingly, the prompt was designed to (i) anchor the assistant as an emotional-support companion, (ii) forbid claims of personal experience, and (iii) discourage clinical positioning, aligning the generated responses with a supportive, non-clinical conversational style.

Beyond the base instruction, two complementary mechanisms to increase consistency and data efficiency were incorporated, as shown in Figure 16. First, few-shot demonstrations to provide in-domain exemplars of the desired interaction pattern (warm acknowledgment, validation, and gentle follow-up), which is particularly relevant given size of ESConv. Because demonstrations can implicitly bias outputs and may lead to evaluation leakage if drawn from held-out data (TOPAL; BOZANTA; BASAR, 2024), the two few-shot examples were sampled exclusively from the validation split and never from the test split.

Second, a persona strategy was applied by injecting a short persona descriptor into the prompt context, encouraging the model to condition its supportive response on user-specific characteristics when available while keeping the underlying objective unchanged. This combined setup—explicit instruction framing, validation-only few-shot demonstrations, and persona conditioning—follows the broader observation that carefully engineered prompts can complement parameter updates Chatterjee et al. (2025). by stabilizing response style and improving instruction adherence in low-resource fine-tuning settings.

4.3 LLM Models

To evaluate the proposed fine-tuning pipeline across model families and capacity regimes, a set of open-weight, instruction-oriented LLMs was selected spanning compact (7–8B) and larger (32B) checkpoints. This selection enables comparison between models that are feasible for iterative experimentation and a higher-capacity alternative that may have a better hability to capture nuanced patterns in emotional-support dialogue.

- ❏ **`Llama2-7B-Chat and Llama-3-8B-Instruct.`** These checkpoints represent two generations of the Llama family, supporting an assessment of whether improvements in base-model quality and instruction alignment yield stronger emotional-support behavior after fine-tuning. In particular, Llama 3 adopts an updated training recipe and enhanced instruction-following alignment, providing a modern reference for dialogue adaptation under instruction prompts (GRATTAFIORI et al., 2024).

- ❏ **`Mistral-7B.`** This model was chosen as a competitive 7B-class alternative with an efficiency-oriented architecture and strong empirical performance under constrained compute and memory budgets. Such design choices make it suitable for parameter-

```
Emotional Support Agent.
Your task is to generate emotional-support responses that are warm,
empathetic, and concise, while respecting strict role constraints.
Role Constraints:
- You are a compassionate AI companion for emotional support.
- You are NOT a human -- never claim personal experiences.
- You are a supportive friend, NOT a clinician.
Few-shot Demonstrations:
Example 1: User: {val_example_1_user}
Assistant: {val_example_1_assistant}
Example 2: User: {val_example_2_user}
Assistant: {val_example_2_assistant}
Current User Message:
{user_message}
Task Description:
Given the user's message, generate a warm, empathetic response that:
1) Acknowledges the user's emotions and validates their feelings.
2) Asks clarifying questions when appropriate.
3) Offers gentle support and practical next steps when appropriate.
4) Avoids clinical language and medical/therapeutic claims.
5) Stays conversational and concise.
Persona Context (optional):
{persona_context}
Answer:
```

Figure 16 – Instruction prompt template used for supervised fine-tuning.

efficient fine-tuning in low-resource settings, complementing the Llama baselines within a comparable parameter scale (JIANG et al., 2023).

- ❑ **DeepSeek-R1-Distill-Qwen-32B.** To complement the 7–8B models, a substantially larger distilled checkpoint was included. Higher model capacity can be advantageous for emotional-support dialogue, where responses often require integrating subtle user cues and maintaining coherence over longer contexts, and distillation-based training aims to preserve instruction-following and reasoning capabilities in an accessible form factor (GUO et al., 2025).

4.4 Fine-Tuning

The fine-tuning stage was conducted using Quantized Low-Rank Adaptation (QLoRA), a parameter-efficient approach that combines low-rank adapters with low-bit quantization of the base model weights (DETTMERS et al., 2023). In QLoRA, the pretrained backbone is kept frozen and stored in a quantized representation to reduce memory footprint, while a small set of trainable low-rank matrices (LoRA adapters) is optimized to specialize the model for the target task. This configuration substantially decreases GPU memory requirements compared to full fine-tuning, while preserving the ability to adapt

instruction-following behavior and domain style (SHI et al., 2025; PATIL; RATHORE; RAMTEKE, 2024). The method is particularly suitable for emotional support applications, where iterative experimentation is required but access to large-scale compute is typically limited.

The adoption of QLoRA is justified by the following considerations:

A. Data privacy and sovereignty. Mental health interactions and associated annotations are highly sensitive and may fall under strict privacy and compliance requirements. Transmitting such data to closed, third-party APIs can introduce avoidable privacy, governance, and compliance risks.

B. Capacity to specialize and mitigate misaligned generic behavior. General-purpose LLMs may respond to mental health scenarios with generic reassurance or overly conservative refusals that can be unhelpful for supportive dialogue. Fine-tuning enables targeted specialization of tone and interaction style (e.g., more structured empathic reflection or motivational interviewing-like phrasing) without retraining the full model from scratch.

C. Hardware accessibility and cost-effectiveness. Full fine-tuning of large models can require multi-GPU setups and high operational cost. QLoRA reduces the memory and compute barrier by keeping the base model quantized and training only a small number of parameters, enabling downstream adaptation with substantially fewer resources than full fine-tuning and improving the overall cost–benefit.

The training configuration followed a standard QLoRA setup with 4-bit quantization using NF4, and LoRA adapters parameterized by rank $r = 16$, scaling factor $\alpha = 32$, and dropout of 0.05. Fine-tuning was performed for 3 epochs. To accommodate GPU memory constraints while maintaining an effective batch size, gradient accumulation was enabled with 16 accumulation steps.

4.4.1 Fine-Tuning Results

This section reports the quantitative outcomes of the fine-tuning experiments under the two preprocessing settings described in section 4.1: (i) the QA-style transformation (single-turn input–output pairs) and (ii) the multi-turn setting obtained via context-aware data splitting. Model performance was evaluated using BERTScore (ZHANG et al., 2020), a reference-based semantic similarity metric that leverages contextual embeddings to compare generated responses against gold references, and has been widely adopted (QIU et al., 2024; YE et al., 2025) for evaluating natural language generation beyond surface-form overlap. For each task configuration, the reported values correspond to the aggregated BERTScore results on the held-out test split.

Table 3 – BERTScore results for fine-tuned models across the QA and multi-turn task settings.

	Task	
Model	**Q/A**	**Multi-turn**
Llama-2-7b	0.8404	0.8321
Mistral-7B	0.8451	0.8422
Llama-3-8B	**0.8562**	**0.8431**
DeepSeek-R1	0.6132	0.7540

Across both evaluation settings, `Llama-3-8B-Instruct` achieved the highest BERTScore, outperforming the alternative checkpoints in the QA formulation (0.8562) and in the multi-turn formulation (0.8431). Therefore, it was selected as the fine-tuned backbone for the subsequent experiments reported in the remainder of this thesis.

Although `DeepSeek-R1` is the largest model evaluated, it achieved the weakest results in both settings. A plausible hypothesis is that this behavior reflects a mismatch between the model's general optimization profile and the target task, which requires short, empathetic, and stylistically constrained emotional-support responses rather than extended reasoning. In addition, under a relatively small domain-specific dataset such as ESConv, larger models may be more sensitive to limited supervision during parameter-efficient fine-tuning, which can hinder effective adaptation. These interpretations remain speculative, but they suggest that model scale alone was not sufficient to guarantee better performance in the present setting.

4.5 Retrieval-Augmented Generation (RAG)

4.5.1 Knowledge Base Document Selection

To ground retrieval-augmented generation in authoritative, evidence-based content, the knowledge base (KB) was constructed from official clinical and public-health guidelines published by the World Health Organization (WHO) and the National Institute for Health and Care Excellence (NICE). This selection strategy prioritizes high-trust sources with explicit recommendations, clear scope statements, and standardized guideline-development processes, which is particularly important in mental health applications where unsupported claims can introduce safety risks. The resulting KB was composed of four core documents, chosen to provide complementary coverage of (1) depression assessment and management, (2) mental health support in non-specialized settings, (3) mental health in occupational contexts, and (4) promotive and preventive interventions for adolescents.

1. **NICE — Depression in adults: treatment and management[1] (NG222).** This guideline consolidates evidence-based recommendations for recognizing, assess-

[1] <https://www.ncbi.nlm.nih.gov/books/NBK583074/>

ing, treating, and managing depression in adults, including stepped-care treatment selection, relapse prevention, and risk considerations. Its inclusion supports retrieval of standardized, high-quality information about depression-related care pathways that can be used to constrain and justify the system's informational content within well-established practice guidance.

2. **WHO — mhGAP Intervention Guide[2] (Version 2.0).** The mhGAP-IG provides structured guidance for the management of priority mental, neurological, and substance use conditions in non-specialized health settings, including depression and self-harm/suicide. It is well-aligned with the intended scope of a supportive chatbot (non-specialist, non-clinical positioning) and enables retrieval of globally oriented, actionable recommendations that emphasize safety, referral, and essential care practices.

3. **WHO — WHO guidelines on mental health at work[3].** These guidelines compile evidence-based recommendations for preventing and addressing mental health risks in workplaces, including organizational interventions, training for managers and workers, and support for return-to-work processes. Given the high prevalence of work-related stressors in mental health narratives, this document strengthens the KB with guidance that is directly relevant to common user contexts while remaining policy- and public-health oriented.

4. **WHO — Guidelines on mental health promotive and preventive interventions for adolescents: Helping adolescents thrive[4].** This guideline provides evidence-informed recommendations for psychosocial interventions targeting adolescents (10–19 years), including universal, targeted, and indicated prevention approaches, with attention to delivery settings such as schools, communities, healthcare, and digital platforms. Its inclusion expands KB coverage to developmentally sensitive content and supports safer, age-appropriate retrieval when user messages involve adolescent-related concerns.

4.5.2 Vector Store Selection

The vector store adopted for the RAG pipeline was chosen based on retrieval effectiveness, latency, and implementation complexity for hybrid retrieval. Initially, ChromaDB[5] was selected due to its lightweight setup, straightforward API, and broad adoption in prototyping retrieval-augmented applications. However, in the targeted setting, ChromaDB did not provide native support for hybrid search retrieval (i.e., combining dense

[2] <https://www.who.int/publications/i/item/9789241549790>
[3] <https://www.who.int/publications/i/item/9789240053052>
[4] <https://www.ncbi.nlm.nih.gov/books/NBK565375/>
[5] Available at: <https://www.trychroma.com/>

vector similarity with lexical signals such as BM25), requiring a manual implementation to approximate hybrid behavior. In practice, this increased engineering overhead and introduced additional latency during retrieval, particularly when scaling beyond small document collections and when repeatedly querying pre-indexed guideline content.

To address these limitations, Elasticsearch[6] was adopted as the vector store for the final system. Elasticsearch provides mature, production-grade indexing primitives and supports hybrid retrieval strategies more directly and efficiently, enabling dense retrieval to be combined with lexical matching in a single retrieval layer. This capability improves retrieval robustness for mental health guidelines, where clinically relevant expressions often depend on exact terminology (lexical match) while user phrasing can be highly variable (semantic match). Consequently, the use of Elasticsearch offered a faster and more effective hybrid retrieval configuration compared to the initial ChromaDB-based prototype, while also reducing the need for custom retrieval logic.

4.5.3 Hybrid Search Retrieval

The RAG pipeline adopts a hybrid search strategy that combines lexical retrieval (BM25) and dense retrieval (embedding-based similarity), followed by rank fusion. This design is supported by evidence that BM25 remains a strong baseline for precise term matching and high-recall retrieval in structured domains (SANNA et al., 2024), while dense embeddings improve semantic generalization when user queries diverge from the vocabulary used in source documents (YANG et al., 2024). In addition, hybrid approaches that integrate both signals have been shown to improve retrieval robustness by leveraging complementary strengths across heterogeneous query formulations (ALJOHANI; ALSANOOSY, 2026).

This combination is particularly well-motivated for mental health applications due to the duality of language in this domain: user messages often mix colloquial expressions, metaphors, and indirect emotional cues with clinically relevant concepts that appear in guidelines under more formal terminology. Dense retrieval helps bridge lexical mismatch by capturing semantic similarity, yet it can be vulnerable to semantic noise, where conceptually opposite statements may become near neighbors because they occur in similar topical contexts (e.g., both appearing in discussions of risk, crisis, or safety). In contrast, BM25 anchors retrieval to explicit terminology and can reduce spurious semantic matches when key terms are present. Therefore, hybrid retrieval improves reliability by combining semantic generalization (embeddings) with lexical precision (BM25), increasing robustness to vocabulary mismatch while mitigating false positives caused by context-driven embedding proximity (ALJOHANI; ALSANOOSY, 2026).

The hybrid retrieval configuration was organized into three layers: (i) indexing and chunking parameters, (ii) retrieval candidate generation, and (iii) fusion via Reciprocal

[6] Available at: <https://www.elastic.co/elasticsearch>

Rank Fusion (RRF). Table 4 summarizes indexing parameters, Table 5 summarizes retrieval parameters, and the RRF scoring function is defined in Eq. 1. Finally, the top-k fused passages (here, $k = 5$) are injected into the LLM prompt as external context.

The lexical and dense retrieval components produce independent relevance scores before fusion. In the lexical branch, BM25 assigns a score to each candidate document based on term matching statistics. In the dense branch, Elasticsearch computes similarity scores using cosine similarity between the query embedding and the document embeddings. These raw scores are not combined directly in the final fusion step. Instead, each retrieval branch produces an ordered ranking list, and the Reciprocal Rank Fusion (RRF) method operates over the rank position of each document in these lists. Therefore, in Eq. 1, $\text{rank}_r(d)$ denotes the position of document d in ranking r, where the ranking itself is induced by the corresponding BM25 or cosine-similarity scores.

Table 4 – Indexing parameters for the hybrid RAG pipeline.

Parameter	Value	Purpose
`Chunk Size`	1000 chars	Controls chunk length extracted from guideline PDFs
`Overlap`	200 chars	Preserves context across adjacent chunks
Embedding model	all-MiniLM-L6-v2	Produces dense representations for semantic retrieval
`Similarity`	cosine	Vector similarity metric in Elasticsearch
`Analyzer`	english	Stemming and stopwords in English

$$\text{score}(d) = \sum_{r \in \text{rankings}} \frac{1}{\texttt{rrf\_k} + \text{rank}_r(d)}. \tag{1}$$

With `rrf_k` $= 60$, a document ranked first in a given list contributes $\frac{1}{62} \approx 0.016$ to the fused score, and documents appearing in both BM25 and dense rankings accumulate contributions. This fusion mechanism balances lexical and semantic retrieval signals without requiring score normalization, yielding a stable top-k set for prompt construction.

Table 5 – Retrieval and fusion parameters for hybrid search.

Parameter	Value	Purpose
`k`	5	Final number of passages returned to the LLM prompt
`rrf_k`	60	RRF constant controlling rank contribution (default)
BM25 candidates	$k \times 2 = 10$	Lexical candidates retrieved before fusion
Dense candidates	$k \times 2 = 10$	Dense candidates retrieved before fusion
`num_candidates` (kNN)	$k \times 2 = 20$	Internal kNN candidate pool in Elasticsearch
BM25 `operator`	`or`	Increases recall by allowing partial term matches

4.6 Guardrails

Guardrails are explicit safety mechanisms designed to constrain the behavior of an AI system under sensitive or high-risk conditions. In the context of mental health conver-

sational agents, their role is to prevent the model from producing unsafe, misleading, or overly unconstrained responses when the input indicates a potential crisis situation. In this work, a guardrail was inserted into the pipeline because the combination of retrieval and free-form generation, although beneficial for grounded and personalized responses, is not sufficient by itself to guarantee safe behavior in all scenarios. Therefore, before any retrieval or generation step is executed, the system applies a dedicated crisis-detection layer to identify inputs that require a deterministic and conservative handling strategy.

A crisis-detection guardrail introduces a highest-priority verification step between preprocessing and retrieval, modifying the inference flow to:

$$\text{Query} \rightarrow \text{Preprocess} \rightarrow \text{Crisis Gate} \rightarrow \begin{cases} \text{Retrieval} \rightarrow \text{Generation}, & \text{if non-crisis} \\ \text{Static Response}, & \text{if crisis} \end{cases}$$

When the crisis gate is triggered, the pipeline is short-circuited: neither document retrieval nor LLM inference is executed, and the system returns a predefined static response. This design reflects the asymmetric cost of false negatives in mental health contexts, prioritizing conservative detection over maximal coverage. The inclusion of explicit safety mechanisms is consistent with prior evidence (SANNA et al., 2024) that health-oriented conversational systems require dedicated safeguards and controlled behaviors to mitigate harmful outputs and improve reliability.

4.6.1 Crisis Detection Architecture

Crisis detection operates through two complementary layers applied to the normalized input text t considering lexical and regular-expression matching, a decision rule and a response mechanism:

- **Lexical matching (keyword matching)**: A controlled vocabulary K_{crise} contains expressions associated with suicidal ideation, self-harm, and explicit planning. The detector activates if at least one keyword is present in the lowercased text:

$$\exists\, k \in K_{\text{crise}} : k \subseteq \text{lower}(t).$$

  In contrast to the broader mental-health relatedness filter, which typically requires multiple keywords or co-occurrence with first-person expressions, the crisis gate uses a unit threshold: a single matched crisis expression is sufficient for activation.

- **Regular-expression matching**: To capture syntactic constructions not reducible to isolated tokens (e.g., “I want to end it all”, “can’t take it anymore”), a set of regular expressions R_{crise} is evaluated:

$$\exists\, r \in R_{\text{crise}} : \text{re.search}(r, \text{lower}(t)) \neq \emptyset.$$

This layer targets passive ideation (e.g., "wish I were dead"), active ideation (e.g., "thinking about killing myself"), and terminal hopelessness cues (e.g., "no reason to live") that may not be reliably captured by keyword lists alone.

- **Decision rule**: The crisis detector returns `True` if either lexical or regex matching fires:

$$\text{is\_crisis\_related}(t) = (\exists\, k \in K_{\text{crise}} : k \subseteq \text{lower}(t)) \;\vee\; (\exists\, r \in R_{\text{crise}} : r \sim \text{lower}(t))\,.$$

- **Response Mechanism**: When `is_crisis_related()` evaluates to `True`, the system returns a deterministic, pre-approved message (static response) and bypasses both retrieval and generation. This ensures predictable behavior under high-risk inputs and prevents uncontrolled content generation in crisis scenarios.

The crisis-detection component relies on a controlled list of crisis-related keywords and regular-expression patterns, which are made available in the GitHub repository associated with this project.[7] When the guardrail is triggered, the system bypasses both retrieval and generation and returns a deterministic static message. The message used in this work is:

Crisis guardrail static response

Your message suggests intense emotional pain, and this is an important moment to seek immediate support. You do not need to face this alone. Please contact the 988 Suicide & Crisis Lifeline now. Call or text 988, available 24 hours a day, 7 days a week.

4.7 Personas

To investigate whether explicit user metadata can induce measurable personalization in model outputs, the experimental setup incorporates personas as structured context injected into the generation prompt and in the query used in retrieval. In this thesis, personas are treated as lightweight user profiles composed of demographic attributes and/or preference descriptors, enabling a controlled analysis of how different types of metadata influence response tailoring in emotional-support dialogue. The underlying hypothesis is that the presence and nature of persona metadata (e.g., clinical background vs. lifestyle preferences) can affect (i) the specificity of supportive suggestions, (ii) the coherence of the assistant's tone with the user profile, and (iii) the consistency of personalization across different inputs.

[7] Available at: <https://github.com/LALIC-UFSCar/AIMHealth-LLM-ConversationalAgent>.

Two complementary persona sources were selected to cover distinct personalization regimes. The first source is AMIVE[8], a Brazilian project situated in a university context, which defines personas representing potential users of their specialized virtual companion framework. The AMIVE personas are characterized by socio-economic and academic attributes (e.g., course and year), technology-use context (e.g., device and preferred communication channels), and, critically, mental-health-related clinical profiles (e.g., anxiety- or depression-related conditions). This design makes AMIVE personas particularly aligned with the target domain of this research, since the metadata directly encodes factors that plausibly shape the form and content of emotional-support responses (e.g., coping constraints, stressors, and preferred interaction style).

The second source is PersonaLens (ZHAO et al., 2025), a benchmark designed for systematic evaluation of personalization in task-oriented conversational assistants. PersonaLens provides user profiles with richer, preference-centric metadata (e.g., interests, dietary constraints, entertainment preferences, travel habits) and interaction-history summaries, reflecting personalization requirements beyond clinical attributes. Although PersonaLens is not specific to mental health, its profiles offer a contrasting metadata structure that emphasizes stable preferences and situational context, enabling an assessment of whether the same personalization mechanisms generalize across heterogeneous profile types.

By combining AMIVE and PersonaLens personas, the experiments compare two distinct forms of personalization signals: (i) domain-aligned, mental-health-centric profiles grounded in a university setting (AMIVE), and (ii) task-oriented, preference-rich profiles intended to probe broader personalization capabilities (PersonaLens). This dual sourcing supports a more rigorous analysis of how different metadata schemas and profile granularity impact the personalization of emotional-support responses, and whether gains are robust across profile types rather than being driven by a single persona design.

Table 6 – Persona sources and metadata characteristics used for personalization analysis.

Source	**Context**	**Personas**	**Representative metadata fields**
AMIVE	University	6	Clinical profile, pain points, therapy status, motivations, academic and social context, technology usage
PersonaLens	Task-oriented benchmark	1.500	Demographics, multi-domain preferences and interests

As illustrated in Figure 17, the Persona-based RAG pipeline combines persona-aware retrieval and generation in a single end-to-end workflow: user input is first screened by a pre-LLM (4.6), and if allowed, persona metadata is extracted to both enrich the retrieval query and condition prompt construction, after which the top-k evidence chunks

[8] <https://amive.ufscar.br/>

retrieved from the indexed knowledge base are injected into the prompt for grounded response generation.

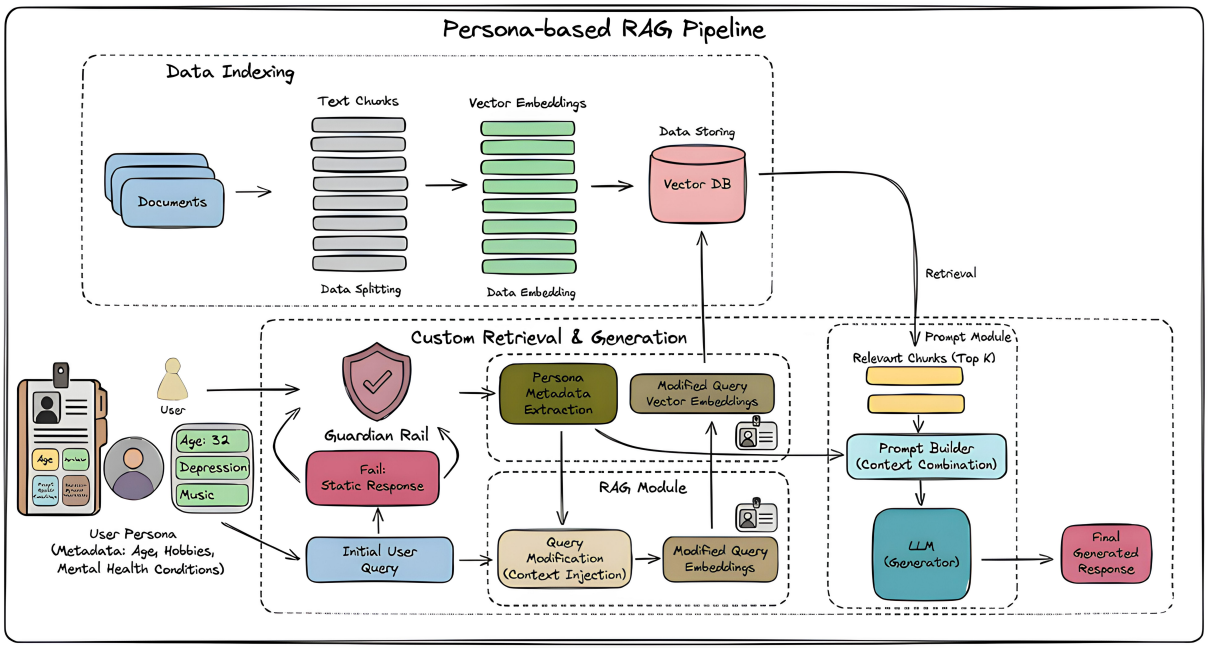

Figure 17 – Pipeline of the PersonaRAG.
This figure was generated with the support of an AI agent Google's Nano Banana 2 (Available at: https://gemini.google.com/)

Chapter 5

Experiments

This chapter describes the experimental protocol used to evaluate personalization and groundedness in emotional-support generation in mental health domain. It covers the construction of three system variants (Baseline, RAG, and Persona-Based RAG), the selection and preparation of input instances for inference, and the two-stage evaluation workflow. The evaluation is organized as (i) a human assessment conducted by domain specialists and (ii) an automated complementary assessment using the LLM-as-a-judge paradigm, enabling both qualitative judgment and scalable comparison across model variants.

The next sections describe the conversational agent variations – baseline (5.1), RAG (5.2) and Persona-Based RAG (5.3) –, the evaluation task (5.4) and protocols – human- (5.5) and LLM-based (5.6).

5.1 Baseline Conversational Agent Construction

The baseline system was built on `Llama-3-8B-Instruct`, selected as the backbone model due to its superior fine-tuning performance among the evaluated checkpoints (see section 4.4.1). The baseline does not use external retrieval or persona metadata; instead, it relies on prompt-based control to shape response quality, safety, and conversational style. In addition to role specification via role-playing instructions (emotional-support companion, non-human, non-clinician), the baseline prompt integrates a structured reasoning scaffold using Chain-of-Thought prompting. Prior work indicates that reasoning-oriented prompting can improve response planning and coherence in generation tasks (YANG et al., 2024). The prompt (Figure 18) therefore includes an explicit stepwise scaffold that encourages the model to identify the user's emotion, validate it, decide whether to explore further, and only then produce the final response.

```
Baseline Reasoning Scaffold (Chain-of-Thought).
Before responding, think through:
1. ACKNOWLEDGE: What emotion is the person sharing?
2. VALIDATE: How can I show this feeling makes sense?
3. EXPLORE: Should I ask questions or offer support?
4. RESPOND: Craft warm, empathetic response
```

Figure 18 – Chain-of-thought scaffold used in the baseline prompt.

In parallel, the baseline applies constraint-based prompting derived from the (ALVES et al., 2023) linguistic guide ("Guia Linguístico"), translated to English for this work[1]. This guide specifies operational do's and don'ts for the chatbot's language to enforce stylistic consistency and a non-clinical stance. The main constraints are summarized below, with concrete examples of compliant (DO) and non-compliant (DON'T) formulations:

❑ **Gender-neutral language.**

- DO: "That sounds really difficult, and it makes sense to feel this way."
- DON'T: "As a man/woman, you should be stronger" or gendered assumptions (e.g., "your boyfriend" without evidence).

❑ **Include non-textual elements sparingly (when appropriate).**

- DO: Use emojis to reduce cognitive load and emphasize the bot's friendliness values
- DON'T: Overuse emojis or use them in heavy moments

❑ **Avoid imperative verbs and overly directive phrasing.**

- DO: "Some people find it helpful to take a short walk or write down what they're feeling."
- DON'T: "Do this now" / "Stop thinking about it" / "You must talk to someone immediately."

❑ **Avoid clinical or service-delivery language.**

- DO: "It sounds like you've been carrying a lot. I can listen if you want to share more."
- DON'T: "Based on your symptoms, this is a diagnosis" / "This requires treatment" / "As your therapist, I recommend...".

[1] Available at:<https://github.com/LALIC-UFSCar/AIMHealth-LLM-ConversationalAgent>

5.2 RAG Conversational Agent Construction

The RAG variant extends the baseline configuration by incorporating external evidence from the mental-health knowledge base while preserving the same backbone model and prompt-control mechanisms. Specifically, the system uses Fine-Tuned-Llama-3-8B-Instruct and retains the Chain-of-Thought scaffold and the AMIVE-based linguistic constraints described for the baseline. The key difference is the addition of retrieved context: for each user input, the hybrid retriever queries the knowledge base and selects the top-k passages (here, $k = 5$) after BM25+dense fusion. These passages are injected into the prompt as an explicit context block that the model can consult when formulating its response.

Providing specialized, retrieval-grounded context has become a common strategy for improving reliability and domain specificity in health-oriented conversational systems, where responses benefit from being grounded in curated, authoritative sources rather than solely in parametric knowledge (ABBASIAN et al., 2024; FAN et al., 2024; VAKAYIL et al., 2024). In addition, retrieval-based augmentation has been adopted to reduce hallucination risk and to increase the relevance of generated content when the user's language is informal, indirect, or partially specified, motivating the use of RAG for mental-health support contexts (YANG et al., 2024). In the present setting, the retrieved passages originate from WHO and NICE guidelines (see section 4.5.1), enabling the model to align its supportive language with evidence-based recommendations and safer informational boundaries.

Operationally, the prompt template follows the baseline structure (role constraints, Chain-of-Thought planning, and linguistic do's/don'ts), but augments it with the retrieved excerpts from the knowledge base. This design aims to improve factual grounding and recommendation consistency by providing authoritative reference material at inference time, reducing reliance on parametric memory alone. By conditioning generation on retrieved guideline content, the RAG variant is expected to maintain a supportive, non-clinical stance while producing responses that are more consistent with validated practices in the mental-health domain.

5.3 Persona-Based RAG Conversational Agent Construction

The Persona-Based RAG variant extends the RAG configuration by incorporating persona metadata at two complementary stages: (i) prompt conditioning context engineering for generation and (ii) persona-aware query reformulation for retrieval. This design follows the general motivation of personalized RAG systems, where user-centric information is leveraged to adapt both retrieval and generation to the user's contextual needs, improving

relevance while reducing noise in retrieved evidence (ZERHOUDI; GRANITZER, 2026a). In parallel, results from personalized conversational information retrieval highlight that the same surface query can correspond to distinct intents across users, and that excessive personalization can introduce query drift and harm retrieval effectiveness; therefore, personalization should be applied selectively and only through relevant profile elements (MO et al., 2025).

- ❑ **Persona injection in the prompt (generation-level personalization).** The baseline prompt structure (role constraints, Chain-of-Thought scaffold, and the AMIVE linguistic guide constraints) is preserved, and the RAG context block (top-k retrieved passages) is maintained. Persona information is added as an explicit Persona Context block placed before the user message and before the retrieved passages (see Figure 17). The prompt instructs the model to consider the persona attributes as contextual signals to tailor tone, examples, and supportive framing, without explicitly repeating sensitive attributes verbatim in the response. This configuration encourages controlled personalization at generation time by making the user profile available as part of the input context, aligning with the broader Persona-Based RAG principle of adapting outputs to user-specific informational needs (ZERHOUDI; GRANITZER, 2026a).

  To better align the inference-time behavior of the model with the objectives of this work, the prompt was designed as a layered instruction template combining role prompting, constraint-based prompting, grounding instructions, persona conditioning, and a reasoning scaffold. This design is intended to ensure that the generated response remains empathetic, contextually relevant, non-clinical, and consistent with the linguistic and behavioral principles established in the Amive project. The full prompt is available at Appendix A.

- ❑ **Persona-aware query reformulation (retrieval-level personalization).** Persona metadata is also used to modify the query sent to Elasticsearch (see Figure 17), aiming to retrieve guideline excerpts that better match the user's latent constraints and context. This follows personalized information retrieval findings that profile-conditioned query reformulation can improve retrieval relevance when personalization is necessary, but should avoid injecting irrelevant personal details that may cause over-personalization and degrade results (MO et al., 2025). Concretely, the system extracts a small set of keywords from the persona profile and performs query expansion (or structured query composition) using three high-salience dimensions:

  - **Age range** (e.g., adolescent vs. adult vs. older adult), used to prioritize age-appropriate guideline content (e.g., adolescent preventive interventions versus adult depression management).

- **Life context** (work, student, retired), used to emphasize context-specific guidance (e.g., workplace recommendations when the persona indicates employment).
- **Pre-existing mental health conditions**, used to bias retrieval toward condition-relevant recommendations and terminology (e.g., depression, anxiety-related conditions), improving robustness when the user's wording is indirect or colloquial.

In practice, this stage operationalizes persona conditioning as lightweight, relevance-oriented query augmentation rather than unrestricted insertion of the full persona text, directly addressing the over-personalization risk emphasized in personalized conversational IR (MO et al., 2025). From a RAG perspective, such user-adapted retrieval is consistent with PersonaRAG-style architectures in which profile signals guide retrieval refinement and ranking to obtain more contextually appropriate evidence for the generator (ZERHOUDI; GRANITZER, 2026a).

5.4 Single-Turn Task

The single-turn task is designed as a one-round conversational experiment to compare the three system variants described in the previous sections (Baseline, RAG, and Persona-Based RAG) under a controlled inference setting. Each input consists of a single user message describing a mental-health-related experience or concern, and each model produces exactly one supportive reply. The objective is to assess which variant produces the best response according to three evaluation criteria: empathy, topic adequacy, and personalization. These criteria are defined and operationalized in section 5.5.

A single-turn protocol was adopted primarily for feasibility: it enables a realistic human evaluation workload by a panel with expertise in human-computer interaction and mental health, while still allowing a complementary scalable assessment via an LLM-as-a-judge stage. In addition, constraining the interaction to one round reduces confounding factors introduced by long context windows (e.g., compounding errors across turns), ensuring that differences among the three variants can be attributed more directly to retrieval grounding and persona conditioning context engineering rather than to multi-turn dialogue management.

5.4.1 Datasets for Single-Turn Inference

To ensure diversity of mental-health narratives and conversational styles, inputs were sampled from four conversational datasets containing user-generated accounts and counseling-oriented exchanges. Together, these datasets cover distinct sources (social media, counseling-

style question answering, and multi-turn counseling) and different topical distributions, supporting a robust comparison of response quality across heterogeneous inputs.

- ❏ Depression Severity Levels Dataset (PRIYADARSHANA; LIANG; PIUMARTA, 2023). This dataset comprises more than 40,000 annotated statements aggregated from multiple social network services and labeled for depression severity rather than binary depression presence. The categorization follows BDI-3 and the Depression Severity Annotation Schema (DSAS), and the dataset was introduced to mitigate data sparsity limitations in severity-level detection by providing a large and comparatively balanced benchmark. In the single-turn setting, these statements provide concise, high-signal user self-reports that require the assistant to respond with appropriate emotional validation while remaining aligned with the topic intensity implied by the severity framing.

- ❏ Reddit Mental Health Dataset (RANI; AHMED; SUBRAMANI, 2024). This dataset is derived from a large-scale Reddit crawl spanning pre-pandemic, during-pandemic, and post-pandemic periods, containing over one million posts from mental-health-related subreddits. A subset was manually annotated to capture root causes and narrative categories, aiming to support the analysis of mental health narratives and their underlying triggers. For the present task, Reddit posts offer long-form user narratives with naturally occurring language, implicit cues, and high variance in detail, demonstrating the robustness of the stress-test to noisy and heterogeneous expressions of distress.

- ❏ MentalChat16K (XU et al., 2025). MentalChat16K is an English benchmark that combines a synthetic counseling dataset with anonymized transcripts of interventions between behavioral health coaches and caregivers in palliative or hospice contexts. The dataset is explicitly motivated by conversational mental health assistance, covering conditions such as depression, anxiety, and grief, and emphasizes privacy-aware curation. Its counseling-oriented structure provides input that resembles help-seeking prompts and enables evaluation of responses that must balance warmth, clarity, and supportive guidance in a way that approximates counseling dialog.

- ❏ Psy-Insight (CHEN et al., 2025). Psy-Insight is an explainable multi-turn bilingual dataset for mental health counseling, constructed from face-to-face multi-turn counseling dialogues and augmented with multi-task annotations and reasoning-oriented explanations. It includes psychotherapy-, emotion-, strategy-, and topic-related annotations, as well as turn-level reasoning and session-level guidance; it also differs from the other sources by providing explicit seeker and supporter roles and being natively multi-turn, with an English portion available for experimentation. In this

thesis, Psy-Insight is included as the only inherently multi-turn source in the single-turn benchmark, and English seeker turns are used to form single-turn inputs while preserving the counseling-specific discourse characteristics.

5.5 Human Evaluation Protocol

Human evaluation follows a standardized manual designed to ensure consistent and comparable judgments across model variants, and the complete manual is publicly available in the project GitHub repository[2]. The task assesses the quality of single-turn emotional-support responses produced by three systems (Baseline, RAG, and Persona-Based RAG) according to three criteria: empathy, topic adequacy, and personalization.

The evaluation manual instructed human judges to assess each response independently and to rely only on the information contained in the user input and associated persona. In this protocol, the three criteria are defined as follows:

- **Empathy**: assesses whether the response acknowledges the user's emotional state in a warm, supportive, and non-judgmental manner. Higher scores indicate that the response shows clear sensitivity to the user's feelings rather than remaining emotionally distant or formulaic.

- **Topic adequacy**: assesses whether the response remains aligned with the user's central concern and provides coherent, relevant support. Higher scores indicate that the response addresses the main issue directly, without drifting away from the topic or introducing irrelevant or incorrect content.

- **Personalization**: assesses whether the response makes appropriate use of details from the user message and persona context, when relevant, without inventing unsupported information. Higher scores indicate that the response is tailored to the specific case rather than being broadly applicable to any user.

Each criterion was rated on a three-point ordinal scale. A score of 1 indicates low quality, corresponding to a response that is inadequate with respect to the criterion (e.g., cold or generic in empathy, off-topic in adequacy, or entirely generic in personalization). A score of 2 indicates intermediate quality, meaning that the response partially satisfies the criterion but remains limited, superficial, or incomplete. A score of 3 indicates high quality, corresponding to a response that clearly satisfies the criterion (e.g., supportive and validating in empathy, directly aligned with the user's concern in adequacy, or specifically tailored to the user and persona context in personalization).

[2] <https://github.com/LALIC-UFSCar/AIMHealth-LLM-ConversationalAgent>

- ❏ **Empathy**: evaluates whether the response acknowledges the user's emotions in a warm, supportive, and non-judgmental manner, demonstrating sensitivity rather than emotional distance or generic reassurance.

- ❏ **Topic adequacy**: evaluates whether the response remains focused on the user's central concern and provides relevant, coherent, and appropriate support without drifting away from the main issue.

- ❏ **Personalization**: evaluates whether the response appropriately incorporates specific details from the user's message and persona context, when relevant, without relying on unsupported assumptions or stereotypes.

Prior to evaluation, the criteria definitions and rating guidelines were reviewed by specialists in human-computer interaction and mental health to improve clarity, reduce ambiguity, and align the rubric with both interactional quality and domain expectations.

Evaluation was conducted through a structured spreadsheet in which each row corresponds to one input instance. In total, the evaluation set contains 24 English inputs associated with 6 personas. For each input, three candidate responses are presented (Response 1/2/3), with the order randomized to blind evaluators to the underlying model identity. For each response, evaluators assign a score from 1 to 3 for each criterion, resulting in 9 ratings per input (3 criteria $\times$ 3 responses). Therefore, each evaluator produces 216 scalar ratings (24 inputs $\times$ 9 ratings), and with 6 evaluators the protocol yields 1,296 ratings in total. When needed, evaluators may add short optional notes (1–3 sentences) to justify scores, flag potential issues (e.g., hallucinated details), or document uncertainty.

To improve annotation quality and reduce fatigue effects, evaluators were instructed to distribute the workload across the full evaluation window (four weeks, from February, 9nd 2026 to March, 6th 2026) instead of completing the task in a single sitting. Several consistency guidelines were enforced:

- ❏ each response had to be scored independently, such that length or politeness alone should not determine the rating;

- ❏ empathy had to be distinguished from agreement, meaning that empathic responses should not reinforce harmful beliefs or risky behavior;

- ❏ persona attributes should only be considered when relevant to the input, so that omission of an irrelevant persona detail should not be penalized; and

- ❏ personalization should not be rewarded when it relies on stereotypes or on inferred attributes that are not explicitly present in the input or persona description.

All texts are evaluated in English; but since all the human judges are native of Portuguese, Portuguese automatic translations are provided only as comprehension support

and should not drive the final scores. In case of uncertainty, evaluators are instructed to choose the more conservative score (lower value) and record a brief justification, reporting unresolved doubts through the support channel while referencing the instance id, dataset, and the response number under discussion.

The evaluator panel comprised two Ph.D.-level researchers, two Ph.D. students, and two undergraduate students, aged between 20 and 35 years. Regarding disciplinary background, four evaluators had training in Psychology and two in Computer Science/Human-Computer Interaction. All evaluators were native speakers of Portuguese; the group included four women and two men, and all identified as White. Although none were native English speakers, all had sufficient proficiency in the language and reported no difficulty assessing English responses.

5.6 LLM-as-a-Judge Evaluation

In addition to the human assessment, an automated evaluation stage was conducted using the LLM-as-a-judge protocol as a complementary counterpoint to expert review. LLM-based judging has been widely adopted as a low-cost and reproducible alternative for ranking open-ended generations, but its validity in expert-knowledge settings remains uncertain and may diverge from specialist judgments, particularly in mental health tasks (SZYMANSKI et al., 2025). For this reason, the automated stage is treated as a comparison signal rather than a replacement for human evaluation: it enables systematic cross-model scoring under the same rubric and, simultaneously, provides an explicit measurement of how well an LLM judge can apply the domain-specific criteria defined in the manual.

The automated protocol follows the same three criteria used in the human rubric (empathy, topic adequacy, and personalization) and applies the same three-point ordinal scale (1–3) for each criterion, as seen in the prompt in B. For each input instance, the judge model receives the user message, the associated persona metadata and the three candidate responses produced by the Baseline, RAG, and Persona-Based RAG variants.[3] The judge is instructed to score each response independently for the three criteria, provide short evidence-grounded rationales based on the text, and then select the best response overall, saving all the results in JSON format.

To reduce bias, candidate responses are presented in randomized order and without revealing which system generated them.

[3] Thus, the LLM judge has the same input as the human judges except from the Portuguese translations.

Chapter 6

Results

This chapter presents the results obtained from the experimental configuration described in chapter 5. The three evaluated systems are referred to as follows: V1 corresponds to the Baseline Conversational Agent, V2 corresponds to the Hybrid Retrieval-Augmented Generation Conversational Agent, and V3 corresponds to the Persona-Based RAG Conversational Agent. Results are reported for both evaluation sources adopted in this work: human evaluation (section 6.1) and LLM-as-a-judge evaluation (section 6.2). For each source, the analysis is organized into three views: overall results, results by persona, and results by evaluation criterion. In all cases, two summary measures are considered: the arithmetic mean of the final scores and the number of wins (best-version selection) obtained by each version.

6.1 Human Evaluation

6.1.1 Overall Results

Table 7 summarizes the overall human evaluation. In this context, *wins* correspond to the number of instances in which a given version achieved the highest final score among the three candidate responses, with the final score (average score) defined as the arithmetic mean of the ratings assigned to empathy, topic adequacy, and personalization. In cases of ties, the win was assigned to all versions sharing the highest final score. According to this measure, V3 obtained the highest number of wins (best-version selection) followed by V2 (51) and V1 (26). In terms of Average Score, V3 also achieved the highest value (2.2593), slightly above V2 (2.2245), while V1 remained substantially lower (2.0671). Therefore, considering both summary measures jointly, the human evaluation indicates V3 as the

best overall version.

Table 7 – Overall human evaluation results.

Metric	V1	V2	V3
Average Score	2.0671	2.2245	**2.2593**
Best-Version Selections	26	51	**67**

6.1.2 Results by Persona

Table 8 presents the human evaluation results by persona.[1] V2 achieved the highest Average Score for Angela, Beatriz, and user51, whereas V3 achieved the highest Average Score for Cauê, user1309, and user228. In terms of Best-Version Selections, V3 was the most frequent winner for Angela, Cauê, user1309, and user228, while V2 led for Beatriz and user51. These results indicate that persona conditioning with context engineering was not uniformly dominant for every profile, but V3 produced the most competitive performance across the full set of personas and concentrated the best gains in four out of the six profiles.

Table 8 – Human evaluation results by persona.

Persona	Metric	V1	V2	V3
Angela	Average Score	2.1111	2.3611	**2.3750**
Angela	Best-Version Selections	4	9	**11**
Beatriz	Average Score	2.0972	**2.3611**	2.1667
Beatriz	Best-Version Selections	0	**13**	11
Cauê	Average Score	2.1250	2.3195	**2.3333**
Cauê	Best-Version Selections	3	8	**13**
user1309	Average Score	2.0833	2.0833	**2.2500**
user1309	Best-Version Selections	7	5	**12**
user228	Average Score	1.9306	1.9722	**2.0972**
user228	Best-Version Selections	5	6	**13**
user51	Average Score	2.0556	**2.2778**	2.1389
user51	Best-Version Selections	7	**10**	7

One plausible explanation for the weaker performance observed in some personas is that persona conditioning may have introduced contextual signals that were not fully coherent with the content of the input message. Because the personas were externally assigned to posts from the evaluation datasets, rather than originating from the same individuals who produced those messages, the model may in some cases have attempted to incorporate persona traits that were not naturally supported by the reported experience. This interpretation is consistent with annotator feedback, as some evaluators reported

[1] The personas *Angela*, *Beatriz*, and *Cauê* were derived from AMIVE, whereas *user51*, *user1309*, and *user228* were derived from PersonaLens.

cases in which personal information from the persona was used incoherently in responses to narratives for which such details were not contextually relevant.

This limitation is likely amplified by the single-turn setting, in which no prior conversational history is available to contextualize or validate the relevance of the injected persona information. As a result, Persona-Based RAG may occasionally produce responses that appear more personalized at the surface level, but less coherent with the user's immediate account, thereby reducing perceived topic adequacy and the overall appropriateness of the response.

6.1.3 Results by Criterion

Table 9 summarizes the human results by criterion. V3 achieved the highest Average Score in Empathy (2.3472) and Personalization (2.1389), while V2 achieved the highest Average Score in Adequacy (2.2569). This pattern indicates that persona conditioning was particularly beneficial for the dimensions most directly related to the intended contribution of the work, namely empathy and personalization, while the RAG component remained especially competitive in topic adequacy when mean score is considered.

This result also suggests an important limitation of the experimental setup. Although persona context engineering improved empathy and personalization, the lower Average Score of V3 in topic adequacy indicates that, in some cases, persona traits were incorporated in a way that was not fully aligned with the content of the input message. According to annotator feedback, some responses introduced user-related details that were available in the persona profile but were not contextually appropriate for the specific message under evaluation. This may have reduced the perceived relevance and coherence of the response with respect to the user's immediate concern. Such behavior is likely related to the design of the experiment itself, since personas were externally assigned to isolated single-turn messages rather than emerging from a multi-turn conversational history. As a result, the persona information was not always naturally grounded in the local context of the interaction, which may have favored occasional mismatches between personalization and topical relevance.

Table 9 – Human evaluation results by criterion.

Criterion	**Metric**	**V1**	**V2**	**V3**
Empathy	Average Score	2.2292	2.3264	**2.3472**
Personalization	Average Score	1.9861	2.1111	**2.1389**
Adequacy	Average Score	1.9861	**2.2569**	2.1875

6.2 LLM-as-a-Judge Evaluation

The evaluation was performed with OpenAI's *GPT-5* and *GPT-5-mini* models during the first half of March 2026, and the outputs are recorded in the same structured format as the human annotation template, enabling direct comparison of score distributions, best-response selections, and agreement patterns between expert ratings and automated judgments.

6.2.1 Overall Results

Table 10 summarizes the overall LLM-as-a-judge evaluation. V3 obtained the highest number of wins (104) and the highest Average Score (2.6619), followed by V1 and V2. Therefore, the LLM-based evaluation indicates V3 as the best overall version. The same overall trend was observed with *GPT-5-mini*. In this case, V3 again obtained the highest number of wins (83) and the highest Average Score (2.4048), while V2 slightly outperformed V1 in Average Score (2.3619 vs. 2.3476). Although the margins were smaller than those observed with *GPT-5*, both evaluator models converged in identifying V3 as the best overall configuration.

Table 10 – Overall LLM-as-a-judge results.

Metric	**Judge**	**V1**	**V2**	**V3**
Average Score	GPT-5	2.5095	2.4452	**2.6619**
Average Score	GPT-5-mini	2.3476	2.3619	**2.4048**
Best-Version Selections	GPT-5	14	22	**104**
Best-Version Selections	GPT-5-mini	23	34	**83**

6.2.2 Results by Persona

Table 11 presents the LLM-as-a-judge results by persona. Under the *GPT-5* evaluation, V3 achieved the highest Average Score for Angela, Beatriz, Cauê, user1309, and user228, whereas V1 achieved the highest Average Score for user51. In terms of Best-Version Selections, V3 was the most frequent winner for all six personas. Under *GPT-5-mini*, V3 again obtained the highest Average Score for Angela, Cauê, user1309, and user228, while V1 remained the highest-scoring version for user51. For Beatriz, V2 and V3 reached the same rounded Average Score (2.4028), with a slight raw-score advantage for V2. In terms of Best-Version Selections, V3 remained the most frequent winner for Angela, Cauê, user1309, user228, and user51, whereas V2 was the most frequent winner for Beatriz. These results suggest that *GPT-5-mini* preserved the same overall tendency toward V3, although with more local variation across personas than *GPT-5* evaluation.

Table 11 – LLM-as-a-judge results by persona.

Persona	Metric	Judge	V1	V2	V3
Angela	Average Score	GPT-5	2.4722	2.5972	**2.7500**
Angela	Average Score	GPT-5-mini	2.2917	2.4722	**2.4861**
Angela	Best-Version Selections	GPT-5	0	4	**20**
Angela	Best-Version Selections	GPT-5-mini	2	6	**16**
Beatriz	Average Score	GPT-5	2.4722	2.4306	**2.6528**
Beatriz	Average Score	GPT-5-mini	2.3611	**2.4028**	2.4028
Beatriz	Best-Version Selections	GPT-5	2	5	**17**
Beatriz	Best-Version Selections	GPT-5-mini	5	**10**	9
Cauê	Average Score	GPT-5	2.4583	2.4444	**2.7222**
Cauê	Average Score	GPT-5-mini	2.3333	2.4444	**2.5417**
Cauê	Best-Version Selections	GPT-5	1	5	**18**
Cauê	Best-Version Selections	GPT-5-mini	2	7	**15**
user1309	Average Score	GPT-5	2.4583	2.4861	**2.6945**
user1309	Average Score	GPT-5-mini	2.3056	2.2778	**2.3195**
user1309	Best-Version Selections	GPT-5	2	3	**19**
user1309	Best-Version Selections	GPT-5-mini	5	4	**15**
user228	Average Score	GPT-5	2.5556	2.2639	**2.5972**
user228	Average Score	GPT-5-mini	2.3195	2.2083	**2.3611**
user228	Best-Version Selections	GPT-5	5	3	**16**
user228	Best-Version Selections	GPT-5-mini	6	5	**13**
user51	Average Score	GPT-5	**2.6667**	2.4500	2.5333
user51	Average Score	GPT-5-mini	**2.4500**	2.4167	2.3000
user51	Best-Version Selections	GPT-5	4	2	**14**
user51	Best-Version Selections	GPT-5-mini	3	7	**10**

6.2.3 Results by Criterion

Table 12 summarizes the LLM-as-a-judge results by criterion. V3 achieved the highest Average Score in Empathy (2.8571) and Personalization (2.3286), while V2 achieved the highest Average Score in Adequacy (2.8500). This pattern is closely aligned with the human evaluation, reinforcing the conclusion that Persona-Based RAG offers the best balance between emotional attunement and personalization, whereas the RAG-only variant remains especially competitive in topic adequacy. When the same analysis was repeated with *GPT-5-mini*, V3 remained the best-performing variant in Personalization (2.1071), and V2 again obtained the highest Average Score in Adequacy (2.6714). However, in Empathy, *GPT-5-mini* assigned a slightly higher Average Score to V1 (2.6143) than to V3 (2.6071). This suggests that criterion-level judgments were somewhat more variable under the smaller evaluator model, even though the overall preference for V3 was preserved.

Table 12 – LLM-as-a-judge results by criterion.

Criterion	Judge	V1	V2	V3
Empathy	GPT-5	2.7500	2.5000	**2.8571**
Empathy	GPT-5-mini	**2.6143**	2.4286	2.6071
Personalization	GPT-5	1.9786	1.9857	**2.3286**
Personalization	GPT-5-mini	1.9286	1.9857	**2.1071**
Adequacy	GPT-5	2.8000	**2.8500**	2.8000
Adequacy	GPT-5-mini	2.5000	**2.6714**	2.5000

6.3 Comparative Analysis

Taken together, the human evaluation and the LLM-as-a-judge results point to the same overall conclusion: V3 is the best version of the personalized mental-health conversational agent developed in this work. In human evaluation, V3 achieved the best overall performance, although V2 remained competitive in specific personas and performed especially well in topic adequacy. In the LLM-based evaluation, V3 also dominated both in Average Score and in Best-Version Selections, suggesting that persona-aware retrieval and prompt conditioning with context engineering contributed positively to the overall response quality. Although the mini variant produced narrower margins and greater local variation across personas and criteria, it also identified V3 as the best overall version in both Average Score and Best-Version Selections.

A relevant point of convergence between human and LLM-based evaluation is that V2 remained especially strong in topic adequacy when mean score is considered, whereas V3 showed stronger gains in personalization-related dimensions. This suggests that retrieval grounding alone already improves the direct alignment of responses with the user's concern, while persona conditioning provides additional benefits in tailoring and emotional attunement. At the same time, the comparison between *GPT-5* and *GPT-5-mini* suggests that criterion-level judgments are somewhat sensitive to the evaluator model, particularly for empathy. This reinforces the role of LLM-as-a-judge as a complementary evaluation signal rather than a substitute for human judgment, especially in dimensions that are more interactional and interpretive in nature.

Overall, these findings support the main argument of this work: combining guideline-grounded retrieval with persona-based conditioning is a promising strategy for improving emotional-support generation in mental health settings. The LLM-as-a-judge experiment further strengthens this conclusion by showing that the preference for Persona-Based RAG is not restricted to a single automated evaluator, but remains visible across different LLM-based judging configurations.

All inputs, generated outputs, and evaluation records produced in this work are publicly available in the project GitHub repository (<https://github.com/LALIC-UFSCar/AIMHealth-LLM-ConversationalAgent>).

Chapter 7

Conclusion

7.1 Conclusion

This work investigated whether the combination of supervised domain adaptation with fine-tuning, retrieval grounding with authoritative mental-health guidelines, and persona-based context engineering can improve the quality and personalization of emotional-support responses generated by large language models. To address this objective, three conversational agent variants were designed and compared under a controlled experimental protocol: a fine-tuned baseline model (V1), a Hybrid Retrieval-Augmented Generation agent (V2), and a Persona-Based RAG agent (V3). The proposed pipeline also incorporated a pre-generation crisis guardrail, so that high-risk inputs could be handled through a deterministic static response rather than unrestricted generation. Within this design, persona information was operationalized as structured metadata and used at two complementary stages of the system, namely retrieval query enrichment and prompt conditioning.

The empirical results support the central hypothesis of this work. In the human evaluation, V3 achieved the best overall performance, with an Average Score of 2.2593 and 67 best-version selections, outperforming V2, which obtained an Average Score of 2.2245 and 51 best-version selections, and V1, which obtained an Average Score of 2.0671 and 26 best-version selections. The same overall tendency was observed in the LLM-as-a-judge evaluation. Under GPT-5, V3 obtained the highest Average Score (2.6619) and the highest number of best-version selections (104), surpassing V1 (2.5095; 14 wins) and V2 (2.4452; 22 wins). This result remained stable under GPT-5-mini, in which V3 again achieved the best overall performance, with an Average Score of 2.4048 and 83 best-version selections, compared with V2 (2.3619; 34 wins) and V1 (2.3476; 23 wins). Taken together, these findings corroborate the hypothesis that the use of persona information as metadata

enriches personalization and improves the overall adequacy and assertiveness of responses with respect to the topic of the posts when integrated into a grounded generation pipeline.

An important implication of these findings is that the contribution of persona information should not be interpreted as a purely stylistic addition to the generated response. In the proposed architecture, persona metadata affects both what is retrieved and how the final answer is constructed. This design allows personalization to influence not only the linguistic form of the response, but also the contextual evidence mobilized to support it. Accordingly, the superiority of V3 in the overall evaluations indicates that the use of persona metadata contributes to a more context-sensitive generation process, capable of producing responses that are not only more personalized, but also more coherent with the user profile and more appropriate to the subject matter expressed in the input message.

From a methodological perspective, this work also contributes a structured framework for studying personalized emotional-support agents under controlled conditions. The study combined fine-tuning, layered prompt design, hybrid retrieval over authoritative mental-health documents, persona-conditioned retrieval, persona-conditioned prompting, and a crisis-detection guardrail within a single experimental pipeline. In addition, the use of both human evaluation and LLM-as-a-judge evaluation provided convergent evidence regarding the superiority of the Persona-Based RAG configuration, while also enabling the analysis of performance by persona and by criterion. In this sense, the main contribution of this work lies both in the empirical evidence that persona metadata improves the overall quality of grounded emotional-support responses and in the methodological framework it offers for future investigations on personalization in sensitive conversational domains.

Although the experimental protocol was conducted in English, the architecture proposed in this work is not inherently language-dependent. Its core components — fine-tuning, hybrid retrieval, persona-conditioned retrieval, persona-conditioned prompting, and guardrail-based control — can be adapted to other linguistic settings, provided that suitable corpora, persona resources, and knowledge-base materials are available. Therefore, the language adopted in the experiments should be understood as a condition imposed by the selected resources and evaluation setup, rather than as a conceptual restriction of the proposed approach.

7.2 Limitations

This work presents important limitations that must be considered when interpreting its results. The first and most significant limitation is the adoption of a single-turn evaluation protocol. Although this design enabled a controlled comparison among the three system variants, it does not fully capture the interactional dynamics of real emotional-support conversations, which typically unfold across multiple turns and depend on progressive disclosure, clarification, memory, conversational continuity, and adaptation over time. As

a consequence, the findings reported here should be interpreted as evidence of response quality at the turn level, rather than as a complete validation of the system in sustained conversational use.

The single-turn design also directly affects the interpretation of the personalization results. Because persona information was associated with isolated posts rather than emerging from an actual interaction history with the same user, the injected persona metadata was not always naturally grounded in the discourse context of the message under analysis. This setting may have reduced the ecological validity of some personalization effects, since real multi-turn interactions would allow persona cues to be progressively validated, refined, and integrated into the dialogue. For this reason, future work should extend the proposed architecture to multi-turn experiments, where it will be possible to evaluate whether persona-conditioned retrieval and generation remain coherent, adaptive, and beneficial across longer conversational trajectories. Such an extension is particularly important for examining properties that cannot be fully assessed in the present setup, including consistency across turns, accumulation of contextual understanding, and the long-term balance between personalization and topical relevance.

A second limitation concerns the experimental language and the scope of the evaluation benchmark. The study was conducted in English due to the characteristics of the selected fine-tuning resources, benchmark datasets, and evaluation protocol. Consequently, the reported findings should not be interpreted as a direct empirical validation for Portuguese or for any other specific language. Nevertheless, this does not restrict the general applicability of the proposed architecture, since its design is independent of the language used in the experiments. The pipeline depends primarily on the availability of appropriate emotional-support data, curated knowledge-base documents, and persona resources that can be operationalized during retrieval and prompting. In addition, the evaluation benchmark remained bounded in size, involving 24 input instances associated with 6 personas and assessed according to three criteria: empathy, adequacy, and personalization. Although these dimensions are directly aligned with the objectives of this work, they do not exhaust all relevant properties of mental-health conversational systems. Future investigations may therefore broaden this framework by incorporating larger and more diverse test sets, multi-turn protocols, longitudinal analyses, and additional human-centered evaluation criteria.

7.3 Future Work

These aforementioned limitations point to directions for future work. First, the evaluation should be expanded to multi-turn interactions, where persona information can emerge more naturally and be validated across conversational history rather than being injected into isolated inputs. Second, future studies should investigate more selective per-

sonalization mechanisms, such as relevance-aware gating strategies that activate persona attributes only when they are clearly supported by the user's message, thus reducing query drift and incoherent tailoring.

It would also be important to broaden the evaluation protocol by using larger and more diverse input sets, a wider pool of external annotators, and additional analyses of agreement and disagreement between expert assessments and LLM-based judgments. Finally, because this thesis is part of the wider AIM-Health project[1] —an international Brazil–UK (United Kingdom) collaboration in AI and NLP for mental health, supported by FAPESP (24/10233-7) and UKRI/MRC — UK Research and Innovation / Medical Research Council — an important next step is to develop and evaluate a Portuguese version of the conversational agent, allowing validation in the linguistic and cultural context most directly targeted by the project.

Altogether, these directions can help establish the robustness, generalizability, and practical applicability of the proposed pipeline in richer conversational and multilingual settings.

[1] <https://www.aim-health.ufscar.br/>

References

ABBASIAN, M. et al. **Conversational Health Agents: A Personalized LLM-Powered Agent Framework**. 2024. Available at: <<https://arxiv.org/abs/2310.02374>>.

AGRAWAL, G. et al. Can knowledge graphs reduce hallucinations in LLMs? : A survey. In: DUH, K.; GOMEZ, H.; BETHARD, S. (Ed.). **Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)**. Mexico City, Mexico: Association for Computational Linguistics, 2024. p. 3947–3960. Available at: <<https://aclanthology.org/2024.naacl-long.219/>>.

ALJOHANI, B.; ALSANOOSY, T. Enhancing medical question answering with llms via a hybrid retrieval-augmented generation framework. **Information**, v. 17, n. 2, 2026. ISSN 2078-2489. Available at: <<https://www.mdpi.com/2078-2489/17/2/133>>.

ALVES, V. de C. et al. College students-in-the-loop for their mental health: a case of ai and humans working together to support well-being. **Interaction Design and Architecture(s) Journal - IxD&A**, n. 59, p. 79–94, 2023. Available at: <<https://doi.org/10.55612/s-5002-059-003>>.

BANERJEE, S.; LAVIE, A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In: GOLDSTEIN, J. et al. (Ed.). **Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization**. Ann Arbor, Michigan: Association for Computational Linguistics, 2005. p. 65–72. Available at: <<https://aclanthology.org/W05-0909/>>.

BENITA, J. et al. Implementation of retrieval-augmented generation (rag) in chatbot systems for enhanced real-time customer support in e-commerce. In: **2024 3rd International Conference on Automation, Computing and Renewable Systems (ICACRS)**. [S.l.: s.n.], 2024. p. 1381–1388.

BORA, A.; CUAYÁHUITL, H. Systematic analysis of retrieval-augmented generation-based llms for medical chatbot applications. **Machine Learning and Knowledge Extraction**, v. 6, n. 4, p. 2355–2374, 2024. ISSN 2504-4990. Available at: <<https://www.mdpi.com/2504-4990/6/4/116>>.

CHAKRABORTY, C. et al. Overview of chatbots with special emphasis on artificial intelligence-enabled chatgpt in medical science. **Frontiers in Artificial Intelligence**, v. 6, 2023. ISSN 2624-8212. Available at: <<https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2023.1237704>>.

CHANG, Y. et al. A survey on evaluation of large language models. **ACM Trans. Intell. Syst. Technol.**, Association for Computing Machinery, New York, NY, USA, v. 15, n. 3, Mar. 2024. ISSN 2157-6904. Available at: <<https://doi.org/10.1145/3641289>>.

CHATTERJEE, A. et al. On the effect of instruction tuning loss on generalization. **Transactions of the Association for Computational Linguistics**, v. 13, p. 1360–1380, 10 2025. ISSN 2307-387X. Available at: <<https://doi.org/10.1162/TACL.a.42>>.

CHEN, K. et al. **Psy-Insight: Explainable Multi-turn Bilingual Dataset for Mental Health Counseling**. 2025. Available at: <<https://arxiv.org/abs/2503.03607>>.

CHO, Y. M. et al. An integrative survey on mental health conversational agents to bridge computer science and medical perspectives. In: BOUAMOR, H.; PINO, J.; BALI, K. (Ed.). **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**. Singapore: Association for Computational Linguistics, 2023. p. 11346–11369. Available at: <<https://aclanthology.org/2023.emnlp-main.698>>.

DAS, A. et al. Conversational bots for psychotherapy: A study of generative transformer models using domain-specific dialogues. In: DEMNER-FUSHMAN, D. et al. (Ed.). **Proceedings of the 21st Workshop on Biomedical Language Processing**. Dublin, Ireland: Association for Computational Linguistics, 2022. p. 285–297. Available at: <<https://aclanthology.org/2022.bionlp-1.27/>>.

DETTMERS, T. et al. Qlora: efficient finetuning of quantized llms. In: **Proceedings of the 37th International Conference on Neural Information Processing Systems**. Red Hook, NY, USA: Curran Associates Inc., 2023. (NIPS '23).

FAN, X. et al. Constructing a knowledge-guided mental health chatbot with LLMs. In: **The 16th Asian Conference on Machine Learning (Conference Track)**. [s.n.], 2024. Available at: <<https://openreview.net/forum?id=FuzY1lFp4V>>.

FITZPATRICK, K. K.; DARCY, A.; VIERHILE, M. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): A randomized controlled trial. **JMIR Ment Health**, v. 4, n. 2, p. e19, Jun 2017. ISSN 2368-7959. Available at: <<http://mental.jmir.org/2017/2/e19/>>.

FU, J. et al. GPTScore: Evaluate as you desire. In: DUH, K.; GOMEZ, H.; BETHARD, S. (Ed.). **Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)**. Mexico City, Mexico: Association for Computational Linguistics, 2024. p. 6556–6576. Available at: <<https://aclanthology.org/2024.naacl-long.365/>>.

GAO, Y. et al. **Retrieval-Augmented Generation for Large Language Models: A Survey**. 2024. Available at: <<https://arxiv.org/abs/2312.10997>>.

GRATTAFIORI, A. et al. **The Llama 3 Herd of Models**. 2024. Available at: <<https://arxiv.org/abs/2407.21783>>.

GUO, D. et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. **Nature**, Springer Science and Business Media LLC, v. 645, n. 8081, p. 633–638, Sep. 2025. ISSN 1476-4687. Available at: <<http://dx.doi.org/10.1038/s41586-025-09422-z>>.

HE, Y. et al. Mental health chatbot for young adults with depressive symptoms during the covid-19 pandemic: Single-blind, three-arm randomized controlled trial. **J Med Internet Res**, v. 24, n. 11, p. e40719, Nov 2022. ISSN 1438-8871. Available at: <<https://www.jmir.org/2022/11/e40719>>.

HUANG, L. et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. **ACM Trans. Inf. Syst.**, Association for Computing Machinery, New York, NY, USA, Nov. 2024. ISSN 1046-8188. Just Accepted. Available at: <<https://doi.org/10.1145/3703155>>.

INKSTER, B.; SARDA, S.; SUBRAMANIAN, V. An empathy-driven, conversational artificial intelligence agent (wysa) for digital mental well-being: Real-world data evaluation mixed-methods study. **JMIR Mhealth Uhealth**, v. 6, n. 11, p. e12106, Nov 2018. ISSN 2291-5222. Available at: <<http://mhealth.jmir.org/2018/11/e12106/>>.

JAFARI, S. et al. Psychological health chatbot, detecting and assisting patients in their path to recovery. In: EL-HAJ, M. (Ed.). **Proceedings of the 1st Workshop on NLP for Languages Using Arabic Script**. Abu Dhabi, UAE: Association for Computational Linguistics, 2025. p. 64–77. Available at: <<https://aclanthology.org/2025.abjadnlp-1.8/>>.

JELINEK, F. et al. Perplexity—a measure of the difficulty of speech recognition tasks. **The Journal of the Acoustical Society of America**, v. 62, n. S1, p. S63–S63, 08 2005. ISSN 0001-4966. Available at: <<https://doi.org/10.1121/1.2016299>>.

JIANG, A. Q. et al. **Mistral 7B**. 2023. Available at: <<https://arxiv.org/abs/2310.06825>>.

KERMANI, A.; PEREZ-ROSAS, V.; METSIS, V. A systematic evaluation of LLM strategies for mental health text analysis: Fine-tuning vs. prompt engineering vs. RAG. In: ZIRIKLY, A. et al. (Ed.). **Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2025)**. Albuquerque, New Mexico: Association for Computational Linguistics, 2025. p. 172–180. ISBN 979-8-89176-226-8. Available at: <<https://aclanthology.org/2025.clpsych-1.14/>>.

LI, H. et al. **Hello Again! LLM-powered Personalized Agent for Long-term Dialogue**. 2024. Available at: <<https://arxiv.org/abs/2406.05925>>.

LIN, C.-Y. ROUGE: A package for automatic evaluation of summaries. In: **Text Summarization Branches Out**. Barcelona, Spain: Association for Computational Linguistics, 2004. p. 74–81. Available at: <<https://aclanthology.org/W04-1013/>>.

LIU, S. et al. Towards emotional support dialog systems. In: ZONG, C. et al. (Ed.). **Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)**. Online:

Association for Computational Linguistics, 2021. p. 3469–3483. Available at: <<https://aclanthology.org/2021.acl-long.269/>>.

_____. **Towards Emotional Support Dialog Systems**. 2021. Available at: <<https://arxiv.org/abs/2106.01144>>.

MAENG, W.; LEE, J. Designing a chatbot for survivors of sexual violence: Exploratory study for hybrid approach combining rule-based chatbot and ml-based chatbot. In: **Proceedings of the Asian CHI Symposium 2021**. New York, NY, USA: Association for Computing Machinery, 2021. (Asian CHI '21), p. 160–166. ISBN 9781450382038. Available at: <<https://doi.org/10.1145/3429360.3468203>>.

MO, F. et al. Towards adaptive personalized conversational information retrieval. In: **Proceedings of the 34th ACM International Conference on Information and Knowledge Management**. New York, NY, USA: Association for Computing Machinery, 2025. (CIKM '25), p. 2137–2147. ISBN 9798400720406. Available at: <<https://doi.org/10.1145/3746252.3761255>>.

PAPINENI, K. et al. Bleu: a method for automatic evaluation of machine translation. In: ISABELLE, P.; CHARNIAK, E.; LIN, D. (Ed.). **Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics**. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, 2002. p. 311–318. Available at: <<https://aclanthology.org/P02-1040/>>.

PATIL, S. S.; RATHORE, A. S.; RAMTEKE, M. Qlora-based fine-tuning of llms on multiple medical reasoning tasks to enhance their comprehension of clinical notes. In: **2024 11th International Conference on Soft Computing Machine Intelligence (ISCMI)**. [S.l.: s.n.], 2024. p. 177–181.

PRIYADARSHANA, Y. H. P. P.; LIANG, Z.; PIUMARTA, I. Heladepdet: A novel multi-class classification model for detecting the severity of human depression. In: TAKADA, H. et al. (Ed.). **Collaboration Technologies and Social Computing**. Cham: Springer Nature Switzerland, 2023. p. 3–18. ISBN 978-3-031-42141-9.

QIU, H. et al. SMILE: Single-turn to multi-turn inclusive language expansion via ChatGPT for mental health support. In: AL-ONAIZAN, Y.; BANSAL, M.; CHEN, Y.-N. (Ed.). **Findings of the Association for Computational Linguistics: EMNLP 2024**. Miami, Florida, USA: Association for Computational Linguistics, 2024. p. 615–636. Available at: <<https://aclanthology.org/2024.findings-emnlp.34/>>.

RANI, S.; AHMED, K.; SUBRAMANI, S. From posts to knowledge: Annotating a pandemic-era reddit dataset to navigate mental health narratives. **Applied Sciences**, v. 14, n. 4, 2024. ISSN 2076-3417. Available at: <<https://www.mdpi.com/2076-3417/14/4/1547>>.

ROBERTSON, S.; ZARAGOZA, H. The probabilistic relevance framework: Bm25 and beyond. **Foundations and Trends® in Information Retrieval**, v. 3, n. 4, p. 333–389, 2009. ISSN 1554-0669. Available at: <<http://dx.doi.org/10.1561/1500000019>>.

SANNA, L. et al. Building certified medical chatbots: Overcoming unstructured data limitations with modular RAG. In: DEMNER-FUSHMAN, D. et al. (Ed.). **Proceedings of the First Workshop on Patient-Oriented Language Processing (CL4Health)**

@ LREC-COLING 2024. Torino, Italia: ELRA and ICCL, 2024. p. 124–130. Available at: <<https://aclanthology.org/2024.cl4health-1.15/>>.

SHI, J. et al. Mentalqlm: A lightweight large language model for mental healthcare based on instruction tuning and dual lora modules. **IEEE Journal of Biomedical and Health Informatics**, p. 1–12, 2025.

SINGH, S.; BENIWAL, H. A survey on near-human conversational agents. **Journal of King Saud University - Computer and Information Sciences**, v. 34, n. 10, Part A, p. 8852–8866, 2022. ISSN 1319-1578. Available at: <<https://www.sciencedirect.com/science/article/pii/S1319157821003001>>.

SOUZA, I.; SOUSA, J. Machado-de. Brazil: World leader in anxiety and depression rates. **Revista Brasileira de Psiquiatria**, v. 39, p. 384–384, 12 2017.

SOUZA, P. de et al. A codesign approach for conversational user interfaces to support college students with depression. In: **Anais do XXIV Simpósio Brasileiro sobre Fatores Humanos em Sistemas Computacionais**. Porto Alegre, RS, Brasil: SBC, 2025. p. 1093–1115. ISSN 0000-0000. Available at: <<https://sol.sbc.org.br/index.php/ihc/article/view/37702>>.

SOUZA, P. M. de et al. Design recommendations for chatbots to support people with depression. In: **Proceedings of the 21st Brazilian Symposium on Human Factors in Computing Systems**. New York, NY, USA: Association for Computing Machinery, 2022. (IHC '22). ISBN 9781450395069. Available at: <<https://doi.org/10.1145/3554364.3559119>>.

SZYMANSKI, A. et al. Limitations of the llm-as-a-judge approach for evaluating llm outputs in expert knowledge tasks. In: **Proceedings of the 30th International Conference on Intelligent User Interfaces**. New York, NY, USA: Association for Computing Machinery, 2025. (IUI '25), p. 952–966. ISBN 9798400713064. Available at: <<https://doi.org/10.1145/3708359.3712091>>.

TOPAL, M. B.; BOZANTA, A.; BASAR, A. Sentiment analysis with llms: Evaluating qlora fine-tuning, instruction strategies, and prompt sensitivity. In: **2024 34th International Conference on Collaborative Advances in Software and COmputiNg (CASCON)**. [S.l.: s.n.], 2024. p. 1–10.

TOUVRON, H. et al. Llama 2: Open foundation and fine-tuned chat models. **ArXiv**, abs/2307.09288, 2023. Available at: <<https://api.semanticscholar.org/CorpusID:259950998>>.

VAKAYIL, S. et al. Rag-based llm chatbot using llama-2. In: **2024 7th International Conference on Devices, Circuits and Systems (ICDCS)**. [S.l.: s.n.], 2024. p. 1–5.

VASWANI, A. et al. Attention is all you need. In: **Proceedings of the 31st International Conference on Neural Information Processing Systems**. Red Hook, NY, USA: Curran Associates Inc., 2017. (NIPS'17), p. 6000–6010. ISBN 9781510860964.

VERESCHAGIN, M. et al. Effectiveness of the minder mobile mental health and substance use intervention for university students: Randomized controlled trial. **J**

Med Internet Res, v. 26, p. e54287, Mar 2024. ISSN 1438-8871. Available at: <<https://www.jmir.org/2024/1/e54287>>.

WEI, J. et al. Chain-of-thought prompting elicits reasoning in large language models. In: KOYEJO, S. et al. (Ed.). **Advances in Neural Information Processing Systems**. Curran Associates, Inc., 2022. v. 35, p. 24824–24837. Available at: <<https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf>>.

XU, J. et al. Mentalchat16k: A benchmark dataset for conversational mental health assistance. In: **Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2**. New York, NY, USA: Association for Computing Machinery, 2025. (KDD '25), p. 5367–5378. ISBN 9798400714542. Available at: <<https://doi.org/10.1145/3711896.3737393>>.

YANG, Z. et al. Chatdiet: Empowering personalized nutrition-oriented food recommender chatbots through an llm-augmented framework. **Smart Health**, v. 32, p. 100465, 2024. ISSN 2352-6483. Available at: <<https://www.sciencedirect.com/science/article/pii/S2352648324000217>>.

YAO, S. et al. Tree of thoughts: deliberate problem solving with large language models. In: **Proceedings of the 37th International Conference on Neural Information Processing Systems**. Red Hook, NY, USA: Curran Associates Inc., 2023. (NIPS '23).

YE, J. et al. SweetieChat: A strategy-enhanced role-playing framework for diverse scenarios handling emotional support agent. In: RAMBOW, O. et al. (Ed.). **Proceedings of the 31st International Conference on Computational Linguistics**. Abu Dhabi, UAE: Association for Computational Linguistics, 2025. p. 4646–4669. Available at: <<https://aclanthology.org/2025.coling-main.312/>>.

YU, H. Q.; MCGUINNESS, S. An experimental study of integrating fine-tuned large language models and prompts for enhancing mental health support chatbot system. **Journal of Medical Artificial Intelligence**, v. 7, n. 0, 2024. ISSN 2617-2496. Available at: <<https://jmai.amegroups.org/article/view/8991>>.

ZERHOUDI, S.; GRANITZER, M. **PersonaRAG: Enhancing Retrieval-Augmented Generation Systems with User-Centric Agents**. 2026. Available at: <<https://arxiv.org/abs/2407.09394>>.

_____. **PersonaRAG: Enhancing Retrieval-Augmented Generation Systems with User-Centric Agents**. 2026. Available at: <<https://arxiv.org/abs/2407.09394>>.

ZHANG, T. et al. Bertscore: Evaluating text generation with bert. In: **International Conference on Learning Representations**. [s.n.], 2020. Available at: <<https://openreview.net/forum?id=SkeHuCVFDr>>.

ZHAO, H. et al. ESC-eval: Evaluating emotion support conversations in large language models. In: AL-ONAIZAN, Y.; BANSAL, M.; CHEN, Y.-N. (Ed.). **Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing**. Miami, Florida, USA: Association for Computational Linguistics, 2024. p. 15785–15810. Available at: <<https://aclanthology.org/2024.emnlp-main.883/>>.

ZHAO, Z. et al. **PersonaLens: A Benchmark for Personalization Evaluation in Conversational AI Assistants**. 2025. Available at: <<https://arxiv.org/abs/2506.09902>>.

ZHENG, C. et al. AugESC: Dialogue augmentation with large language models for emotional support conversation. In: ROGERS, A.; BOYD-GRABER, J.; OKAZAKI, N. (Ed.). **Findings of the Association for Computational Linguistics: ACL 2023**. Toronto, Canada: Association for Computational Linguistics, 2023. p. 1552–1568. Available at: <<https://aclanthology.org/2023.findings-acl.99/>>.

Appendix

APPENDIX A

Persona-Based RAG Agent Prompt

As described below, the final inference prompt was designed as a layered instruction template combining role constraints, linguistic constraints, grounding rules, personalization rules, and a silent reasoning scaffold. The complete prompt is presented in Box A.

Optimized Inference Prompt

```
Emotional Support Agent.
Your task is to generate emotional-support responses that are warm, empathetic,
and concise, while respecting strict role constraints.
IDENTITY AND SCOPE
- You are not a human and must never claim personal experiences, feelings,
memories, or real-world actions.
- You are not a clinician, therapist, psychiatrist, psychologist, or diagnostic
system.
- You do not diagnose, label, or prescribe treatment.
- Your role is to offer emotional support, gentle guidance, and a safe space for
expression.

PRIMARY GOAL
Generate a supportive response that is:
- empathetic,
- relevant to the user's concern,
- personalized when user context is available,
- grounded in the retrieved reference information when relevant,
- non-clinical, natural, and conversational.

STYLE AND LINGUISTIC CONSTRAINTS
Gender-neutral language.
- DO: "That sounds really difficult, and it makes sense to feel this way."
```

```
- DON’T: "As a man/woman, you should be stronger" or gendered assumptions (e.g.,
"your boyfriend" without evidence).
- Include non-textual elements sparingly (when appropriate).
- DO: Use emojis to reduce cognitive load and emphasize the bot’s friendliness
values.
- DON’T: Overuse emojis or use them in heavy moments.
Avoid imperative verbs and overly directive phrasing.
- DO: "Some people find it helpful to take a short walk or write down what they’re
feeling."
- DON’T: "Do this now" / "Stop thinking about it" / "You must talk to someone
immediately."
Avoid clinical or service-delivery language.
- DO: "It sounds like you’ve been carrying a lot.  I can listen if you want to
share more."
- DON’T: "Based on your symptoms, this is a diagnosis" / "This requires treatment"
/ "As your therapist, I recommend...".

GROUNDING RULES
- If retrieved reference information is relevant, use it to improve the response.
- Do not fabricate facts, recommendations, or resources that are not supported by
the retrieved information or the user context.
- If persona information is available, use it to adapt tone, examples, and focus.
- Never explicitly reveal that you are using persona metadata or retrieved
context.

PERSONALIZATION RULES
If persona information is available, adapt the response using relevant signals
such as:
- age range,
- life context (e.g., student, worker, retired),
- emotional patterns,
- ongoing difficulties,
- motivations or goals,
- pre-existing mental health conditions.
Only use persona details when they are relevant to the current message.
Do not infer new personal attributes beyond what is provided.

INTERNAL RESPONSE SCAFFOLD
Before answering, silently follow this structure:
1.  ACKNOWLEDGE: identify the main emotion or distress signal in the user’s
message.
2.  VALIDATE: express why the feeling makes sense without exaggeration or
over-identification.
3.  EXPLORE: decide whether a gentle follow-up question or a small supportive
suggestion is more appropriate.
4.  RESPOND: write one warm, concise, emotionally supportive response.

OUTPUT CONSTRAINTS
- Write a single response only.
```

```
- Do not use bullet points.
- Do not mention these instructions.
- Do not output the reasoning steps.

OPTIONAL PERSONA CONTEXT
{persona_context}

REFERENCE INFORMATION
{retrieved_context}

USER MESSAGE
{user_message}
```

APPENDIX B

LLM-as-Judge Evaluation Prompts

As described in Section 5.6, the LLM-as-a-judge evaluation pipeline was implemented using two prompts: a system prompt, which defines the evaluation rubric and scoring constraints, and a user prompt template, which provides the persona, user input, and candidate responses for each evaluation instance. The complete prompts are presented in Box B.

LLM-as-Judge System Prompt

```
You are an expert evaluator of mental-health chatbot responses.
Your task is to evaluate chatbot responses according to a standardised rubric.
You must evaluate EACH response INDEPENDENTLY. Do NOT compare responses with each
other.
Scores must be integers from 1 to 3.  If uncertain between two scores, choose the
LOWER one.

=== EVALUATION RUBRIC ===

EMPATHY
1 (Low) = Cold, generic, unwelcoming or slightly judgmental; barely considers the
user's emotions.
2 (Medium) = Considers the user's emotions and tries to be welcoming, but in a
superficial, standardised or unspecific way.
3 (High) = Clearly considers the user's emotions, is warm, caring and
non-judgmental; shows genuine interest and sensitivity.

THEMATIC ADEQUACY
1 (Low) = Off-topic, changes subject, distorts the problem, or brings
wrong/irrelevant information.
2 (Medium) = Partially aligned with the topic:  recognises the main subject but
drifts away or does not elaborate adequately.
```

```
3 (High) = Directly focuses on the central topic and offers coherent
interventions/suggestions compatible with the topic.

PERSONALIZATION
1 (Low) = Generic:  would serve any person; ignores details from the user's
message and/or persona.
2 (Medium) = Uses 1-2 details from the user's situation (or persona when
relevant), but still sounds partially generic.
3 (High) = Tailored:  connects several specific details from the user's message
and context (including persona when relevant), without inventing information.  The
response seems written for this specific user.

=== IMPORTANT RULES ===
- Evaluate each response independently (do not compare them).
- Being empathetic does NOT mean reinforcing harmful beliefs or validating risky
behaviors.
- Do NOT penalise a response for not using all persona characteristics when they
are not relevant; DO penalise when it ignores essential characteristics from the
user's message.
- Do NOT give a high Personalization score if the response assumes beliefs/values
the user has not mentioned or that are not in the persona description.
- Consider the persona context and the input message in your evaluation.
- The evaluation is based on the ENGLISH text of the responses.
- Return your evaluation as a JSON object with EXACTLY this structure (no
additional text):

{
  "response_1": {
    "empathy": <int 1-3>,
    "thematic_adequacy": <int 1-3>,
    "personalization": <int 1-3>
  },
  "response_2": {
    "empathy": <int 1-3>,
    "thematic_adequacy": <int 1-3>,
    "personalization": <int 1-3>
  },
  "response_3": {
    "empathy": <int 1-3>,
    "thematic_adequacy": <int 1-3>,
    "personalization": <int 1-3>
  }
}
```

LLM-as-Judge User Prompt Template

```
=== PERSONA ===
Name/ID: {persona_name}
```

```
{persona_description}

=== USER INPUT (message to evaluate responses against) ===
{user_input}

=== RESPONSE 1 ===
{response_1}

=== RESPONSE 2 ===
{response_2}

=== RESPONSE 3 ===
{response_3}

Evaluate each response independently using the rubric. Return ONLY the JSON object.
```