2024 11th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)
Amity University, Noida, India. Mar 14-15, 2024

Mental Healthcare Chatbot based on custom diagnosis documents using a quantized Large Language Model

Ayush Kumar
Department of Computer Science and Engineering
KIET Group of Institutions
Ghaziabad, India
ayush.2024cse1170@kiet.edu

Sanidhya Sharma
Department of Computer Science and Engineering
KIET Group of Institutions
Ghaziabad, India
sanidhya.2024cse1120@kiet.edu

Shreyansh Gupta
Department of Computer Science and Engineering
KIET Group of Institutions
Ghaziabad, India
shreyansh.2024cse1150@kiet.edu

Dharmendra Kumar
Department of Computer Science and Engineering
KIET Group of Institutions
Ghaziabad, India
dharmendra.kumar@kiet.edu

***Abstract*— This research presents a novel retrieval-based question-answering (QA) framework utilizing LangChain's (version 0.1.6) modular architecture and Chainlit's (version 0.7.700) conversational interface. Our system efficiently handles PDF and directory documents, generates sentence embeddings with HuggingFace's pre-trained model, stores vectors in FAISS for fast search, employs the powerful CTransformers (version 0.2.27) with Llama-2-7B-Chat-GGUF model, and guides it with a custom prompt template for accurate and factual responses. The integrated Chainlit interface facilitates user interaction, demonstrating the framework's potential for knowledge-intensive domains like medical chatbots.**

Keywords*— *retrieval-based QA, CTransformers, sentence embedding, chatbot, large language models (LLMs), quantization, vector store

I. INTRODUCTION

The prevalence of mental health illnesses is extensive, however, numerous persons encounter challenges in obtaining therapy as a result of various structural impediments, such as restricted accessibility, exorbitant expenses, and the social disgrace linked to requesting assistance. Nevertheless, the progress in digital technology, the availability of the Internet, and widespread ownership of smartphones can overcome these obstacles by providing anonymity, scalability, and cost-efficiency. Empirical digital interventions, such as online platforms and mobile applications, have demonstrated their efficacy. Nevertheless, they frequently encounter issues with little user involvement and inadequate compliance. By integrating proactive human assistance into these digital technologies, adherence rates, and outcomes can be enhanced. However, this method restricts the potential to scale up. A potential way to address this obstacle could involve the creation of conversational agents, commonly known as chatbots. These chatbots replicate human conversation using written language, which might potentially improve user engagement and facilitate automation for scalability. This could be especially beneficial in low- and middle-income settings characterized by a substantial disparity in mental health care and limited accessible resources.

Exploring the vast and ever-growing body of mental health research presents a significant challenge for those seeking to extract meaningful insights, we have developed a novel question-answering (QA) system specifically designed to effectively navigate PDF-based mental health research documents. Our system harnesses the power of vector databases and the quantized Llama 2 language model to deliver informative answers to user queries.

We present a framework tailored for resource-constrained settings, especially focusing on local deployments with limited CPU and memory availability. Llama 2 is an open-sourced family of generative text Large Language Models (LLMs) by Meta. We use the llama-2-7b.Q5_K_M.gguf and similar .ggml models, which are chat-optimized, 7 billion parameters, quantized ~4.7GB models.

A. Large Language Models (LLMs)

LLMs, or Large Language Models, are a specific kind of artificial intelligence (AI) that undergo extensive training using vast quantities of textual data. The data can encompass several forms of content, such as books, essays, code, and even social media posts. The training procedure facilitates the acquisition of patterns and associations among words by LLMs, enabling them to execute many tasks, including text generation, text summarization, and question answering.

B. Quantization

Quantization is a method used to decrease the size and computational expenses of Large Language Models (LLMs) by minimizing the amount of memory required to store parameters. It involves converting model parameters (weights and activations) from high-precision floating-point numbers (typically 32-bits and 16-bits) to lower-precision data types (usually 4-bits). There are two common approaches used to quantize models:

1) Quantization-Aware Training: In this approach, we induce "fake quantization" operations in the model's training graph, allowing the model to adapt to reduced precision while potentially preserving accuracy.

2) Post-Quantization Training: A pre-trained model is converted to a lower precision in this approach. This allows for a simple setup and fast training. Nevertheless, this trade-off results in reduced adaptability and precision.

For our purpose, we chose to go with the latter approach as we wanted a technically non-rigorous solution that was easy to set up on a local device with limited compute resources (majorly CPU).

C. Vector Storage

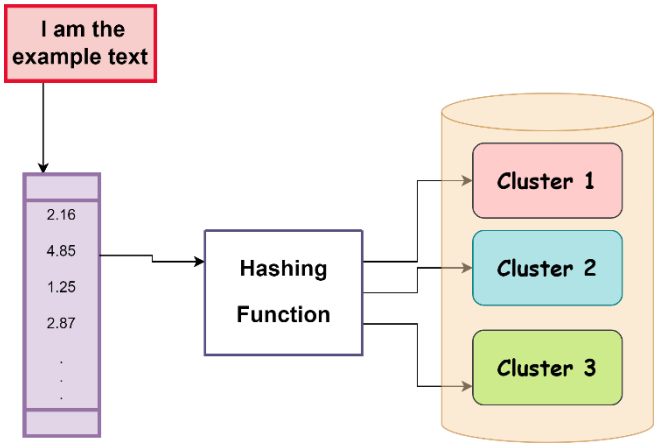

Fig.1. Illustration showing how vectors are stored in a vector database

Vector databases are a specialized type of database designed to store and efficiently retrieve data represented as high-dimensional vectors. These vectors are mathematical representations of data objects, capturing their characteristics and relationships. Unlike traditional relational databases that rely on keywords and exact matches, vector databases excel at similarity search.

Vectors are ordered lists of numbers that represent multidimensional data points (for plain text in this case). These vectors are stored in different clusters according to their similarity (using techniques like Cosine similarity, Jaccard index, etc.) determined by the hashing function in the vector database. When a query is performed, the plaintext is converted into Query Vector which is pointed to a cluster by the hashing function. From there, the most relevant stored vectors (similar to the query) are retrieved. This is the basis of the document-based QA retrieval system. A huge amount of text from a document is divided into chunks which are converted into vectors and stored in the database.

D. CTransformers

CTransformers is a library that provides Python bindings for Transformer models implemented in C/C++ using the GGUF library. Its core function is to enable leveraging the speed and efficiency of C/C++ implementations of advanced natural language processing models within Python projects.

E. Retrieval-Augmented Generation (RAG)

Retrieval-augmented generation has emerged as a powerful paradigm within natural language processing (NLP), leveraging the complementary strengths of retrieval-based and generative models to enhance the quality and factual consistency of the generated text.

Retrieval models play a crucial role in unearthing relevant information from vast troves of data. Utilizing techniques like information retrieval and semantic search, they sift through document collections or knowledge bases to identify the most pertinent content aligned with a specific query. While retrieval models excel at precision and factual accuracy, their inherent strength lies in relocating existing information, rendering them incapable of generating novel or imaginative content.

Conversely, generative models are engineered to produce novel content in response to a provided prompt or context. Language models, which utilize vast amounts of training data, acquire knowledge of the structures and patterns present in natural language. While generative models can generate texts that are imaginative and logically consistent, they might encounter difficulties when it comes to empirical precision or contextual relevance.

This methodology incorporates the capability of retrieval (or browsing) into the generation of LLM texts. The system integrates a retriever system, responsible for extracting pertinent document excerpts from a vast corpus, with an LLM, which generates responses utilizing the data contained within those excerpts. RAG enables the model to essentially "look up" external information to enhance its responses [2].

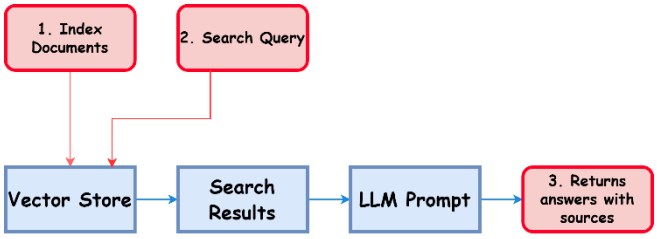

Fig.2. Primitive Architecture of RAG

II. Related Work

Traditional language representation models, like ELMo and GPT, were limited by their unidirectional nature. This meant they could only access information from either the left or right context of a word, which restricted their ability to understand complex relationships in language. BERT addresses this limitation by introducing a novel bidirectional pre-training approach. By joint conditioning on both left and right context in all layers of its Transformer-based architecture, BERT can learn deeper and more comprehensive representations of language, opening new possibilities for improved performance on complex NLP tasks [1].

In general, quality assurance systems that function in an open domain follow a two-stage pipeline. The first step is passage retrieval, which involves extracting relevant text segments from a knowledge base that are relevant to the input query. TF-IDF and BM25 were examples of sparse vector techniques that were previously employed in the retrieval of documents. On the other hand, in recent times, researchers have shifted towards the implementation of dense text representations, which enable a more profound understanding of textual similarity in terms of semantics. An example of such an implementation can be found in the DPR (Dual Encoder with Passage Ranking) model, which combines two BERT models to produce embeddings for text passages and queries [3]. By taking the dot product of these embeddings, the similarity of the passage to the query is

represented. It has been demonstrated by DPR that improving retrieval precision increases QA precision overall. The application of extractive language models such as BERT and generative language models like BART/GPT-2 in the context of answer generation has been the subject of recent research [4].

The training of language models with quantization awareness faces various challenges such as slow training, high memory usage, and concerns regarding data privacy when accessing the complete training set. Conversely, post-training quantization leads to faster training but compromises adaptability and precision. To address these issues, a method called module-wise reconstruction error minimization (MREM) was introduced. This approach utilizes a small subset of training data at a time to minimize the layer-wise reconstruction error caused by quantization. Optimizing all the interconnected layers within each module allows for larger granularity and achieves significantly improved results with minimal memory overhead [5].

Raymond et al. studies highlight the challenges in digital healthcare transformations stating that people may face the baby duck syndrome, sticking to conventional methods out of habit. However, the significant advantages of advanced systems like 24/7 support, personalization, and privacy may help them coexist with current methods [6]. A similar review suggests that the use of clinical practice guidelines during training virtual mental health agents is essential to make them explainable and safe [7].

Direct preference optimization (DPO) has emerged as a promising technique for aligning large language models (LLMs) with human preferences, offering a simpler alternative to reinforcement learning (RL)-based methods. Unlike RL, which relies on complex reward functions and exploration-exploitation trade-offs, DPO directly leverages human-labeled preference pairs {(prompt, preferred response, unpreferred response)} to fine-tune the LLM. This simplicity translates to efficient training and ease of implementation, making it attractive for various LLM-related research. However, challenges remain. DPO can be susceptible to overfitting on limited preference data, requiring strategies like regularization (e.g., Identity Preference Optimization) to improve generalizability. Additionally, incorporating nuanced preference types (e.g., factual correctness, creativity) into the DPO framework remains an active area of research. Despite these challenges, DPO's effectiveness and relative ease of use make it a valuable tool for exploring LLM alignment and preference-guided generation within LLM-RAG research [13].

According to Balcombe et al.'s literature review, AI chatbots that offer consistent and pertinent assistance for anxiety among university students and depression and hyperactivity in adults have the potential to produce favorable clinical outcomes [8]. A preliminary evaluation of Vitalk, a mental health chatbot in Brazil, has demonstrated the potential of chatbots in reducing anxiety, depression, and stress symptoms [9]. The study by Dosovitsky and Bunge specifically focuses on the development of a chatbot for depression in adolescents, emphasizing positive user experiences [10]. Another study conducted on college students using the Tess chatbot concluded that while there was no significant effect on depressive symptoms, the initial results showed that Tess was effective in addressing anxiety symptoms [11]. A Chinese study introducing Emohaa, a conversational agent offering mental health support through CBT-guided exercises and emotional venting options, suggests significant reductions in depression, negative affect, and insomnia symptoms after using Emohaa. Additionally, the generative dialogue platform improved long-term insomnia, highlighting the promise of such agents in future mental health support [14].

Scoping reviews, like the one conducted by Ahmed et al., suggest that chatbots have a promising future, especially in areas where one-on-one psychiatrist-patient conversations are not viable [12]. An independent samples t-test revealed a substantial beneficial impact of chatbot emotional disclosure on user satisfaction and desire to reuse a chatbot counseling service, following social penetration theory, uncertainty reduction theory, and the CASA framework. Hayes (2013) found that user emotional disclosure intention and perceived intimacy with a chatbot counselor have a mediating role in the connection between chatbot emotional disclosure and user satisfaction and intention to reuse the chatbot counseling service in Model 6. Additionally, we discovered a sequential mediation of the user's purpose to disclose emotions and perceived closeness with a chatbot counselor in the connection between emotional disclosure by the chatbot and user pleasure and intention to reuse the chatbot counseling service [15].

III. Research Motivation and Problem Statement

Imagine a mental health companion, readily available, personalized, and knowledgeable, guiding you through difficult times. This is the vision driving our research behind a customizable, AI-powered chatbot tailored for mental health support. Traditional mental health resources often face limitations: limited access, long wait times, and cost constraints. Even when available, accessing support can be daunting due to stigma and the emotional vulnerability involved. The model aims to bridge this gap by offering readily accessible, confidential, and non-judgmental support.

Our motivations stem from a deep understanding of the challenges individuals face in accessing mental health care. We observe:

- *Limited Availability:* Traditional therapy appointments are often booked weeks or months in advance, leaving individuals struggling to find immediate support.
- *Cost Barriers:* Therapy expenses can be prohibitive, especially for vulnerable populations.
- *Stigma and Accessibility:* The stigma surrounding mental health and geographic limitations can further hinder access to care.
- *Need for Personalization:* Individual needs and circumstances vary greatly, requiring personalized support tailored to their specific situations.

The proposed model addresses these challenges through several key innovations:

- *AI-powered assistance:* Utilizing the power of large language models, the solution offers 24/7 support, readily available whenever needed.
- *Privacy and Security:* User privacy and data security are paramount. The system operates under

strict ethical guidelines and secure data storage protocols.

- *Personalization:* The system adapts to individual needs and preferences, offering customized guidance and support.
- *Accessibility and Cost-effectiveness:* The proposed work aims to be an accessible solution to everyone, regardless of location or financial constraints.

IV. Proposed Methodology

For building the basic model, two broad approaches were considered: one being fine-tuning an open-sourced LLM for our use case, and the other being using a semantic search-enriched question-answering system. After careful consideration of the principles, advantages, and disadvantages involved, we decided to move forward with the latter approach because of the following reasons:

Broader Knowledge Coverage: Semantic search with LLM question-answering employs a dual-step methodology wherein it initially identifies pertinent excerpts from an extensive collection of documents and subsequently formulates responses based on those excerpts. This approach enables the provision of more precise and current information by harnessing the latest data from diverse sources. Conversely, fine-tuning LLM relies on the knowledge embedded in the model during training, which can potentially become obsolete or insufficient as time progresses.

Context-focussed Responses: It can produce more accurate and targeted responses by utilizing specific information from relevant documents. On the other hand, fine-tuned models may generate answers that rely on the general knowledge stored within the model, resulting in less precise or unrelated responses to the given question.

Flexibility: The semantic search component offers the flexibility to effortlessly incorporate fresh information sources or customize it for various domains, enhancing its adaptability to specific use cases or industries. Conversely, refining necessitates retraining the model, a process that can consume significant time and computational resources.

Efficient Ambiguous Query Handling: It can clarify queries by pinpointing the most pertinent passages associated with the question. This can result in more precise and pertinent responses, in contrast to a well-adjusted LLM that might encounter difficulties in dealing with ambiguity in the absence of appropriate context.

Hardware Limitations: The solution is targeted at common people, who usually do not have the time, hardware resources, or technical proficiency required to fine-tune an LLM. It would be much easier for them to upload their diagnosis documents and reports.

A. Ingestion Phase

In the ingestion phase, the user uploads the (diagnosis recommendations, reports, etc.) document(s) in a designated directory. The text is then divided into numerous chunks. These chunks are then converted into text embeddings using a pre-trained sentence-transformers model using the host device's computation resources (CPU). The chunks and their corresponding embeddings are then stored in a vector database for efficient similarity search and retrieval.

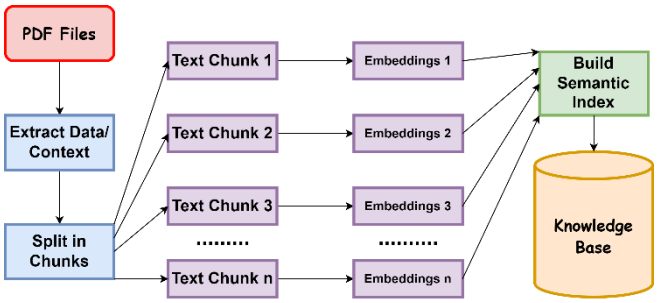

Fig.3. Ingestion Phase

B. Question-Answering Phase

The user will input their question on the interface, which will be converted into text embeddings using the same sentence-transformers model that was used to generate the embeddings of text chunks. A semantic search using the resulting vectors will be performed in the vector database. The most appropriate text chunks will be ranked in order and then the LLM will generate an answer based on them.

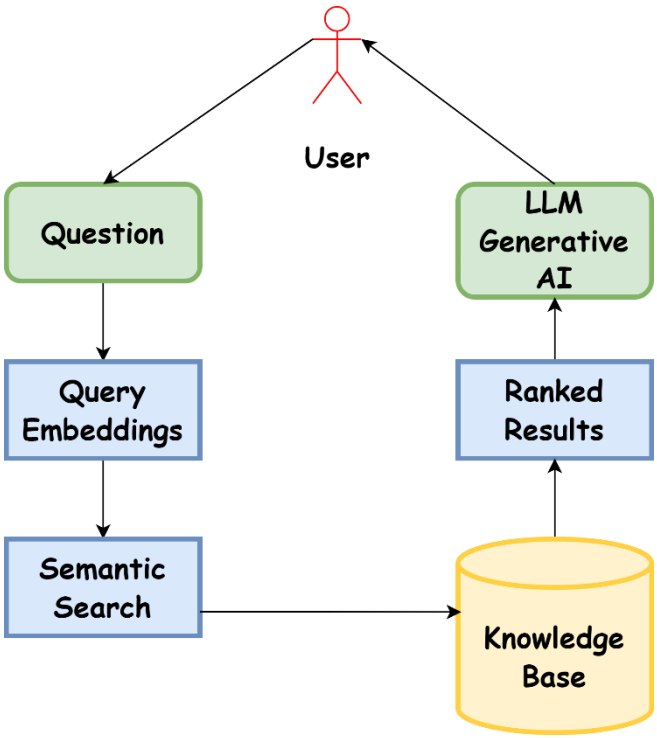

Fig.4. Generation of responses for user query

V. Model Architecture

The solution required an open-sourced LLM that could be quantized to run on the common computer hardware. The quantized Llama 2 .ggml and .gguf models were used for this purpose. The original Llama 2 model was trained on publically available online data resources.

Mental health conversations and question-answer pairs-related datasets on Kaggle and sample Medical Summary Reports available for informational use from SOAR providers on the Substance Abuse and Mental Health Services Administration website were used to compile multiple 10MB and 45MB documents to measure the effectiveness of the chatbot.

FAISS was used to create the vector database that could store embeddings and chunks on the local memory. It is an open-source library that provides efficient similarity search and clustering algorithms for high-dimensional vectors.

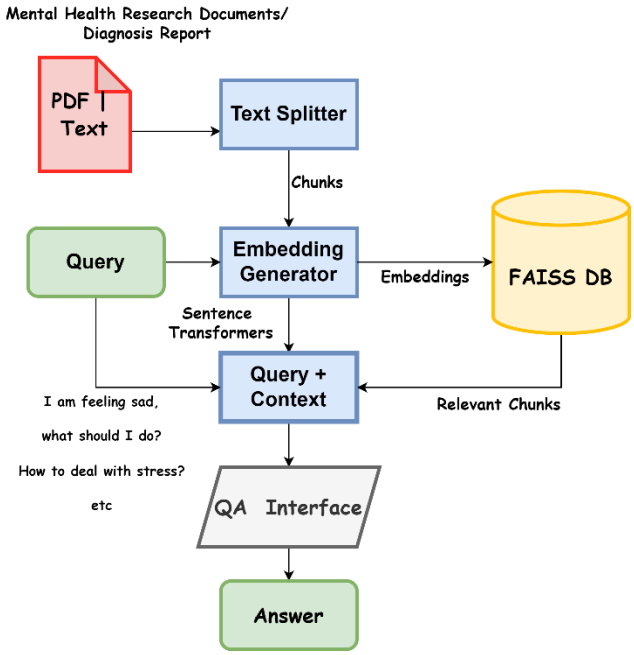

Fig.5. Overview of Model Architecture

Chainlit was used to make the interface. It is an open-source async Python framework that makes it incredibly fast to build AI applications. We can visualize the "chain of thought" and debug the queries and prompts on the interface, which makes it a good fit for our solution.

Fig.6. "Chain of thought" on the interface

VI. Observations

The prototype was run through multiple quantized Llama 2 7B models on a standard 8GB RAM computer. The responses were checked with the source file uploaded and marked as low/medium/high following whether the response was accurate (as per source data) and satisfactory or not.

TABLE I. QUALITY OF ANSWERS BY MODELS

Model Name (ggml/gguf)	Bits	Size (GB)	Max. RAM(GB)	Answer Quality (~10MB file)	Answer Quality (~45MB file)
ggmlv3. q3_K_L	3	3.60	6.10	Low	Low
ggmlv3. q4_K_M	4	4.08	6.58	Medium	Low
ggmlv3. q5_1	5	5.06	7.60	High	Medium
Q3_K_M. gguf	3	3.30	5.80	Low	Low
Q4_K_M. gguf	4	4.08	6.58	Medium	Low
Q5_K_M. gguf	5	4.78	7.28	High	Medium

- For a k-bit (k=3,4,5…) quantized model, the GGUF version uses less RAM than the GGML version.
- The GGML and GGUF versions did not exhibit significant differences concerning the quality of the responses they produced.
- The quality of answers improved drastically as the k-bit (5>4>3) quantization increased.
- 'Strict' prompts like "…Just say 'I don't know' if the query cannot be answered accurately using source text. Do not make things up." significantly reduced hallucinations.

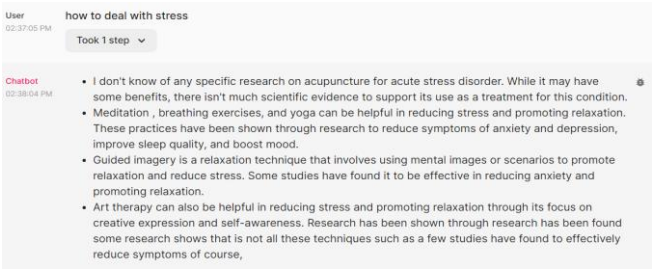

Fig.7. Sample Response of the 4-bit quantized GGML model

VII. Conclusion And Future Scope

The research aimed to study the feasibility and effectiveness of using post-training quantized models in generating responses to user queries based on uploaded documents. This solution will serve as a basic architecture for analysis and further improvements. The quality of the generated responses can be improved by fine-tuning the model on a basic set of healthcare domain knowledge, such as BioMistral - which utilizes Mistral as the foundational model further pre-trained on PubMed Central data [16], which would serve as a better alternative to directly using the base quantized Llama 2 7B Chat model. Implementing the Module-wise Quantization Error Minimization (MREM) strategy can alleviate the challenges associated with the substantial memory overhead and sluggish training that occurs during the post-training quantization of models. As discussed, if a fixed model for responding to queries of a narrow domain is required, quantization-aware training on the domain corpus will be the preferable option.

We focussed on a solution that could be used on common hardware by users without having the knowledge and technical proficiency regarding large language models. Bigger models will give better-quality answers in general. This may be verified by using GPT-4 benchmark evaluation. Using question-answer pairs created from the corpus for training, and/or employing a model that utilizes Direct

Preference Optimization - like Zephyr 7B, can significantly improve performance.

Using GGUF models is preferred as we can offload some layers to GPU if it is available. The study suggests the usage of at least 5-bit quantized models for practical use cases.

By examining the integration of established methods with MLC LLM, a universal solution that enables the deployment of language models across a wide range of backend hardware and native applications, it is possible to create mobile applications that incorporate these chatbots, thus increasing their accessibility to the users.

ACKNOWLEDGEMENT

We express our sincere gratitude to Tom Jobbins (known as The Bloke on HF) for his invaluable contributions to the field of Large Language Models through the provision of quantized LLMs. His active involvement has significantly enhanced the landscape of LLM development, and we appreciate his dedication to advancing this critical area of research.

REFERENCES

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", 2019.

[2] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela, "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks", 2020.

[3] Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih, "Dense Passage Retrieval for Open-Domain Question Answering", 2020.

[4] Shamane Siriwardhana, Rivindu Weerasekera, Elliott Wen, Tharindu Kaluarachchi, Rajib Rana, and Suranga Nanayakkara, "Improving the Domain Adaptation of Retrieval Augmented Generation (RAG) Models for Open Domain Question Answering", 2023.

[5] Haoli Bai, Lu Hou, Lifeng Shang, Xin Jiang, Irwin King, and Michael R. Lyu, "Towards Efficient Post-training Quantization of Pre-trained Language Models", 2022.

[6] Raymond R. Bond, Maurice D. Mulvenna, Courtney Potts, Siobhan O'Neill, Edel Ennis, and John Torous, "Digital Transformation of Mental Health Services", 2023.

[7] Surjodeep Sarkar, Manas Gaur, Lujie Karen Chen, Muskan Garg, and Biplav Srivastava, "A Review of the Explainability and Safety of Conversational Agents for Mental Health to Identify Avenues for Improvement", 2023.

[8] Luke Balcombe, "AI Chatbots in Digital Mental Health", 2023.

[9] Kate Daley, Ines Hungerbuehler, Kate Cavanagh, Heloísa Garcia Claro, Paul Alan Swinton, and Michael Kapps, "Preliminary Evaluation of the Engagement and Effectiveness of a Mental Health Chatbot", 2020.

[10] Gilly Dosovitsky and Eduardo Bunge, "Development of a Chatbot for Depression: Adolescent Perceptions and Recommendations", 2020.

[11] Maria Carolina Klos, Milagros Escoredo, Angela Joerin, Viviana Noemí Lemos, Michiel Rauws, and Eduardo L. Bunge, "Artificial Intelligence–Based Chatbot for Anxiety and Depression in University Students: Pilot Randomized Controlled Trial", 2021.

[12] Arfan Ahmed, Asmaa Hassan, Sarah Aziz, Alaa A Abd-alrazaq, Nashva Ali, Mahmood Alzubaidi, Dena Al-Thani, Bushra Elhusein, Mohamed Ali Siddig, Maram Ahmed, and Mowafa Househ, "Chatbot Features for Anxiety and Depression: A Scoping Review", 2023.

[13] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn, "Direct Preference Optimization: Your Language Model is Secretly a Reward Model", 2023.

[14] Sabour, Zhang, Xiao, Zhang, Zheng, Wen, Zhao, and Huang, "A chatbot for mental health support: exploring the impact of Emohaa on reducing mental distress in China", 2023

[15] Park, Chung, and Lee, "Effect of AI chatbot emotional disclosure on user satisfaction and reuse intention for mental health counseling: a serial mediation model", 2022.

[16] Labrak, Bazoge, Morin, Gourraud, Rouvier, and Dufour, "BioMistral: A Collection of Open-Source Pretrained Large Language Models for Medical Domains", 2024.