Virtual Therapy using Large Language Models and Retrieval Augmented Generation

1st Prof. Amruta Paul
Computer Science and Engineering Yeshwantrao Chavan College of Engineering
Nagpur, India
amrutatapas@gmail.com

2nd Vivek Karadbhajne
Computer Science and Engineering (AIML)
Yeshwantrao Chavan College of Engineering
Nagpur, India
vivekkaradbhajne23@gmail.com

3rd Arnab Chakraborthy
Computer Science and Engineering (AIML)
Yeshwantrao Chavan College of Engineering
Nagpur, India
mailarnab@gmail.com

4th Jayushna Mahadule
Computer Science and Engineering (AIML)
Yeshwantrao Chavan College of Engineering
Nagur, India
mailjayushna@gmail.com

5th Sharwari Hatwar
Computer Science and Engineering (AIML)
Yeshwantrao Chavan College of Engineering
Nagpur, India
mail.sharwari@gmail.com

6th Aryan Kadbe
Computer Science and Engineering (AIML)
Yeshwantrao Chavan College of Engineering
Nagpur, India
mail.aryan@gmail.com

Abstract:
Mental health issues such as anxiety, depression, and panic attacks are increasing among individuals particularly in age group 20-35. Due to social stigma and costly therapy with long waiting queues, many avoid seeking help, which results in worsening conditions over time. To address this challenge, we propose an expert system leveraging Retrieval-Augmented Generation (RAG) and transformer-based fine-tuned large language models to create a virtual counselling chatbot. This system is designed to be easily accessible, empathetic and provide context-aware responses similarly like a professional therapist.

Keywords:

Mental Health, Virtual Therapy, RAG, LLM, Chatbot, Depression, AI Therapist, Chatbot.

I. Introduction

There is a significant rise in mental health-related problems in young adults aged 20-35 [1][2].Changes in life due to academic, professional and financial issues combined with lack of awareness about mental health leads to fewer people seeking help. Social stigma and negative views regarding mental illness makes things even worse, resulting in people allowing minor illnesses to escalate to a severe level.

Despite the availability of effective treatment and research in this domain, such research is limited to standard books and experienced counsellors. Which results in a substantial treatment gap, particularly in low- and middle-income countries, where more than 75% of individuals in this age group suffer psychological disorders. Barriers such as a lack of awareness and the high cost of counselling further discourage people.

According to WHO psychological interventions like counselling and therapy can be highly effective for many mental health conditions, particularly depression, panic, and anxiety. These interventions can even be delivered by trained non-specialists, including community workers, volunteers, and even parents.

Recent AI advancements [3][4][5] in dialogue systems and LLMs for psychological counselling have been successful to some extent in generating helpful and engaging back-and-forth interactions with clients to heal themselves; however, there is still a major problem with these chatbots, like hallucination [6], where these models deviate from the context and provide irrelevant responses, which may confuse the patients.

To address this issue, we propose a RAG-based architecture with fine-tuned LLM designed to provide contextually accurate & empathetic Reponses. [7]

II. Background And Related Work

Recent development in Large-Language models and introduction to techniques like RAG can be utilized to address mental health related issues. This section explores prior work in mental health domain and research done to resolve such issues.

1. Self-Evaluation Using Questionnaires

Health questionnaires, specifically designed for mental health severity evaluation helps to represent the severiety in numeric results. These results helps in self-diagnosis and continuous monitoring of depression and anxiety levels. Regular data collection helps identify patterns and trends that can be exploited to provide therapy [8]. An application called CareBot uses PHQ assessments to offere anonymous counselling to Indian students[10]. Additionally, LLMs such as ChatGPT have been tested with specialized questionnaires like the SCPQ (Stress and Coping Process Questionnaire) [11] to improve their understanding of emotional contexts through staged scenarios.

Although these questionnaires provide a valuable method for assessing a patient's mental state, the effective representation and automation of this data remain areas of ongoing research.

2. Vector Databases

Data representation in vector forms has gained significant momentum due to its efficient and scalable applications in AI. Vector databases are categorised into two types:

Specialised and generalised vector databases. Specialised vector databases are explicitly designed and optimised for storing and querying vector data, whereas generalised vector databases support vector data management within a relational database.[13].A specialized vector database offers offer quick and precise nearest-neighbour search, clustering, and similarity matching. Due to their scalability and high performance, they are more preferred by AI applications [14]. Various data embedding methods transforms data from formats such as text, images, and audio into numerical representations, known as embeddings. These embeddings offer several benefits, including fast search capabilities, scalable architecture, and improved efficiency for AI applications.[15] Despite these advantages, Vector databases face challenges when handling massive datasets for tasks like similarity search. Optimizing performance in such scenarios requires advanced techniques, including enhanced indexing querying methods, and efficient similarity search algorithms. [16]

3. Hallucination in LLMs

Hallucinations in large language models (LLMs) can be described as produced content that is either irrational or untrue to the original material.. These hallucinations are categorized into two types:

- Intrinsic Hallucinations: Where the output contradicts the source content.
- Extrinsic Hallucinations: Where the generated output cannot be verified against the source.

Hallucination generally arises due to issues in data, training, and inference stages Common problems include poor quality, misinformation, bias, and outdated knowledge[17][18]. To address these issues, various strategies have been developed, particularly focusing on fine-tuning LLMs with domain-specific data which is factually correct and free from bias or misinformation. The computational burden of traditional fine-tuning techniques can be mitigated using methods like Low-Rank Adaptation (LoRA). LoRA optimizes fine-tuning by reducing trainable parameters and memory requirements[20].

Retrieval Augmented Generation (RAG) integrates external knowledge sources to improve the accuracy. This method reduces reliance on the LLM alone, which improves generalization on out-of-domain tasks while minimizing resource requirements.[19]

4. LLM & RAG

A combination of RAG and fine-tuned LLM know as Retrieval-Augmented Fine-Tuning (RAFT) have been developed. RAFT is a post-training recipe that enhances LLMs for domain-specific RAG tasks. In RAFT, the model is trained to distinguish relevant documents from distractors and retrieve most useful information. RAFT uses a chain-of-thought-style reasoning process, which improves the model's ability to reason and deliver accurate responses. [29]

The combination of RAG and LLM bridges the gap between intrinsic model knowledge and external, dynamic knowledge base. [23].

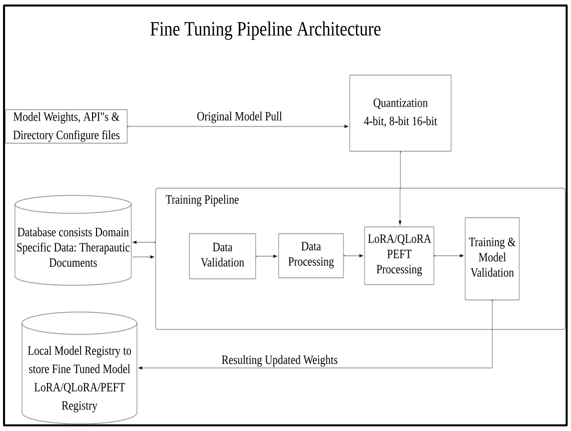

Fine Tuning Pipeline Architecture
Model Weights, API"s & Directory Configure files
Original Model Pull
Quantization 4-bit, 8-bit 16-bit
Training Pipeline
Database consists Domain Specific Data: Therapautic Documents
Data Validation
Data Processing
LoRA/QLoRA PEFT Processing
Training & Model Validation
Local Model Registry to store Fine Tuned Model LoRA/QLoRA/PEFT Registry
Resulting Updated Weights

Figure 1 Fine Tuning Architecture

RAFT separates the retrieval and generation processes, where we can trace the generated outputs back to the sources. This improves trust in the system and enables verification. [24]

The openness of RAFT makes it a must choice for medical domain-related tasks.

The RAFT's performance further can be improved by incorporating prompt engineering techniques like few-shot prompting. Such practise makes the models achieve more accuracy in question answering or conversation-related tasks, outperforming larger models despite having significantly fewer parameters. [26][28]

III. Research GAP

Despite significant advancements in understanding mental health issues and increasing accessibility to traditional therapy, there remains a notable gap in leveraging virtual counseling to address mental health challenges. With recent advancements in AI technology and RAFT technique we can improve the current results and develop an expert system which could potentially more personalized, and effective mental health support

IV. Methodology

We developed a virtual counselling system to address the counselling needs faced by young individuals. The system leverages large language fine-tuned models (LLMs) integrated with retrieval-augmented generation (RAG) to provide scalable, accessible, and interactive therapy session through chat.

1) Data Collection

The foundational data for the RAG application were collected from books and structured questionnaires:

a) Books:

Three standard psychology books (*Book 1*, *Book 2*, and *Book 3*) were selected for their widespread use in educational institutions and credibility among counsellors. These books are trusted resources for psychological theories, practices, and case studies.

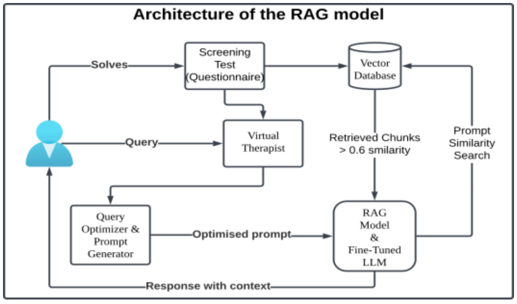
b) Questionnaires

The application integrates widely accepted mental health evaluation tools such as PHQ-9 (Patient Health Questionnaire) and GAD-7 (Generalised Anxiety Disorder Assessment). These tools assess users' mental states and disease severity.

2) Challenges In Preprocessing

The data gathered was in the format of PDF documents which often contains introductory sections such as author details, forewords, and publisher information, which are irrelevant for our use. To address this issue, the following steps were implemented:

1. **Manual Analysis**: Each PDF was reviewed to identify and exclude unnecessary sections. For instance, if the first X pages of a document were deemed irrelevant, they were excluded.
2. **Page Number Parameterisation:** A page number parameter was incorporated during chunking. Chunks retrieved with page numbers corresponding to discarded sections were excluded from further processing.
3. **Negative Indexing:** To improve the process performance, pages that were identified as irrelevant were indexed negatively. If a retrieved chunk carried a negative page number, it was automatically discarded.
4. **Vector Store:** After preprocessing and chunking the data, a vector store was initialised to efficiently manage and retrieve relevant information. Each
5. chunk was padded with metadata to improve traceability and retrieval relevance.
6. **Meta Data:** Book Name, Page Number, Character Count, Word Count, Sentence Count, Token Count, Number of Chunks


Architecture of the RAG model
Solves
Screening Test (Questionnaire)
Vector Database
Query
Virtual Therapist
Retrieved Chunks > 0.6 smilarity
Prompt Similarity Search
Query Optimizer & Prompt Generator
Optimised prompt
RAG Model & Fine-Tuned LLM
Response with context

During vector store initialisation, the chunk text and its metadata were embedded. These metadata help us improve the searching accuracy and is essential for proving validation of data, which we can easily do from the metadata.

3) Dynamic Model Selection

Due to computational constraints, a dynamic selection mechanism was developed to optimise the use of advanced models like Gemini and GPT-3.5 Turbo even with less computational power. The quantisation level of the models is dynamically adjusted based on the available GPU or CPU computational power. For resource-constrained environments, quantised versions of the models are employed without compromising the quality of responses.

For users requiring access to state-of-the-art models, full versions of Gemini and GPT-3.5 Turbo can be deployed on a cloud platform, accessible via the system's website.

4) Implementation Of Retrieval-Augmented Generation (Rag)

The RAG architecture integrates retrieval mechanisms with generative capabilities to minimise hallucinations and enhance response reliability. RAG

a. Query Preprocessing:

The input query is subjected to syntactic and semantic analysis to detect errors, ambiguities, or inconsistencies. A query refinement module restructures the input into a semantically robust format, which helps identify the most relevant chunks from the vector store.

b. Document Retrieval:

The refined query is passed to the retrieval module, which calculates the similarity of the query with the document chunks stored in the vector store using cosine similarity or analogous metrics; the vector store evaluates the relevance of each chunk against the query.

c. Similarity Threshold:

After extensive experimentation across datasets, a similarity score threshold of 0.6 was finalised. Chunks with score meeting or exceeding this threshold are retrieved and passed as context to the LLM.

5) Contextualised Response Generation:

The retrieved chunks are appended as the context for the LLM to generate semantically coherent and factually accurate responses. In addition to data, a prompt engineering technique known as few-shot prompting was used to make sure the responses are engaging. LLM uses both intrinsic knowledge and the provided context to deliver personalized, engaging, factually correct, and precise outputs.

Our virtual counselling system represents a comprehensive mental health solution aimed at addressing the growing need for accessible and interactive therapy. The proposed system leverages a retrieval-augmented generation (RAG) architecture integrated with large language models (LLMs) to provide personalized, context-aware responses. The features of this system include the use of vector databases for efficient document retrieval, dynamic model selection based on computational resources, and secure storage of user interactions to ensure privacy. Developed with a robust preprocessing pipeline and advanced prompt engineering, the system utilises trusted psychology books and standardized mental health questionnaires as foundational data sources, making it a reliable tool for early mental health interventions.

V. Results

Our virtual counselling system is designed using retrieval-augmented generation (RAG) to deliver context-aware, personalized therapeutic interactions. efficiently retrieves relevant chunks of data based on the user's input query. In our experimental setup with random question setups for counselling, acting like a patient, we found that the RAG model achieved an accuracy of over 85% in retrieving contextually relevant documents and generating coherent responses. The system also incorporates large language models (LLMs), dynamically selected based on available computational resources, to provide real-time, empathetic responses even when not connected to the internet. The response generation takes not more than 3 s, which is ideal for a smooth and timely interaction. This rapid response time contributes to a positive user experience and improves user engagement. To enhance therapeutic accuracy and patient understanding, the system integrates standard mental health questionnaires like PHQ-9 and GAD-7, which help assess users' mental states. We observed that users found these evaluations highly effective in helping them identify their mental health status early in life.

For system evaluation, we used several Key Performance Indicators (KPIs) such as response accuracy, relevance, user engagement, and overall satisfaction. The proposed system successfully maintains high relevance in generated responses. Furthermore, user feedback indicated that the therapy sessions feel personal, helpful, and supportive, with an average response time of 2.5 seconds. Real-world testing revealed that the system efficiently handles various user queries, accurately processing input and delivering responses aligned with the user's emotional needs. In addition, the interactions are anonymized, ensuring that sensitive user data is securely stored and not shared with any third parties. The integration of a cloud-based deployment for access to state-of-the-art models without quantisation, like Gemini and GPT-3.5 Turbo, provides the most in-depth therapeutic responses.

VI. Conclusion

Overall, our virtual counselling system has demonstrated substantial success in providing early-stage mental health support, with proven capabilities in emotional understanding, privacy protection, and user satisfaction. The system's ability to combine powerful language models with retrieval mechanisms ensures that it is both effective and scalable for diverse user needs and makes it accountable.

VII. References

[1] Islam, M. R., Rahman, M. S., & Qusar, M. S. (2022). Community-based decentralized mental health services are essential to prevent the epidemic turn of post-Covid mental disorders in Bangladesh: A call to action. Health Science Reports, 5(4). https://doi.org/10.1002/hsr2.734

[2] Rosales-Ricardo, Yury, et al. "Prevalence of burnout syndrome in university students: A systematic review." *Salud mental* 44.2 (2021): 91-102.

[3] June M Liu, Donghao Li, He Cao, Tianhe Ren, Zeyi Liao, and Jiamin Wu. 2023. Chatcounselor: A large language models for mental health support. *arXiv preprint arXiv:2309.15461* (2023).

[4] Huachuan Qiu, Hongliang He, Shuai Zhang, Anqi Li, and Zhenzhong Lan. 2023a. Smile: Single-turn to multi-turn inclusive language expansion via chatgpt for mental health support. *arXiv preprint arXiv:2305.00450* (2023).

[5] Huachuan Qiu, Tong Zhao, Anqi Li, Shuai Zhang, Hongliang He, and Zhenzhong Lan. 2023b. A benchmark for understanding dialogue safety in mental health support. In *CCF International Conference on Natural Language Processing and Chinese Computing*. Springer, 1–13.

[6] W. de Almeida da Silva, L. C. Costa Fonseca, S. Labidi and J. C. Lima Pacheco, "Mitigation of Hallucinations in Language Models in Education: A New Approach of Comparative and Cross-Verification," *2024 IEEE International Conference on Advanced Learning Technologies (ICALT)*, Nicosia, North Cyprus, Cyprus, 2024, pp. 207-209, doi: 10.1109/ICALT61570.2024.00066.

[7] L. Shi, M. Kazda, B. Sears, N. Shropshire and R. Puri, "Ask-EDA: A Design Assistant Empowered by LLM, Hybrid RAG and Abbreviation De-hallucination," *2024 IEEE LLM Aided Design Workshop (LAD)*, San Jose, CA, USA, 2024, pp. 1-5, doi: 10.1109/LAD62341.2024.10691824.

[8] S. Lee, S. Lee and Y. Lee, "Self-Monitoring the Mental-Health State of a Focused Population with Multiple Self-Questionnaires and Sentiment Descriptions," *2024 IEEE 7th International Conference on Multimedia Information Processing and Retrieval (MIPR)*, San Jose, CA, USA, 2024, pp. 674-680, doi: 10.1109/MIPR62202.2024.00116.

[9] P. Singh, K. K. Srinivas, A. Peddi, B. Shabarinath, I. Neelima and K. A. Bhagavathi, "Artificial Intelligence based Early Detection and Timely Diagnosis of Mental Illness - A Review," 2022 International Mobile and Embedded Technology Conference (MECON), Noida, India, 2022, pp. 282-286, doi: 10.1109/MECON53876.2022.9752219

[10] R. Crasto, L. Dias, D. Miranda and D. Kayande, "CareBot: A Mental Health ChatBot," 2021 2nd International Conference for Emerging Technology (INCET), Belagavi, India, 2021, pp.15, doi:10.1109/INCET51464.2021.9456326.

[11] R. Crasto, L. Dias, D. Miranda and D. Kayande, "CareBot: A Mental Health ChatBot," 2021 2nd International Conference for Emerging Technology (INCET), Belagavi, India, 2021, pp.15, doi:10.1109/INCET51464.2021.9456326.

[12] Limbachia, Janak, Yash Damani, Shubh Dave and Vidya Sagvekar. "MOODIFY: Tailored, Personal and Multifaceted AI Assistant for Young Adult Mental Health Issues." *2023 6th International Conference on Advances in Science and Technology (ICAST)* (2023): 106-110.

[13] Y. Zhang, S. Liu and J. Wang, "Are There Fundamental Limitations in Supporting Vector Data Management in Relational Databases? A Case Study of PostgreSQL," 2024 IEEE 40th International Conference on Data Engineering (ICDE), Utrecht, Netherlands, 2024, pp. 3640-3653, doi: 10.1109/ICDE60146.2024.00280.

[14] P. N. Singh, S. Talasila and S. V. Banakar, "Analyzing Embedding Models for Embedding Vectors in Vector Databases," *2023 IEEE International Conference on ICT in Business Industry & Government (ICTBIG)*, Indore, India, 2023, pp. 1-7, doi: 10.1109/ICTBIG59752.2023.10455990

[15] S. Kukreja, T. Kumar, V. Bharate, A. Purohit, A. Dasgupta and D. Guha, "Performance Evaluation of Vector Embeddings with Retrieval-Augmented Generation," *2024 9th International Conference on Computer and Communication Systems (ICCCS)*, Xi'an, China, 2024, pp. 333-340, doi: 10.1109/ICCCS61882.2024.10603291.

[16] X. Xie, H. Liu, W. Hou and H. Huang, "A Brief Survey of Vector Databases," *2023 9th International Conference on Big Data and Information Analytics (BigDIA)*, Haikou, China, 2023, pp. 364-371, doi: 10.1109/BigDIA60676.2023.10429609

[17] Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., & Liu, T. (2024). A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Transactions on Information Systems*. Association for Computing Machinery (ACM). DOI: 10.1145/3703155.

[18] Xu, Z., Jain, S., & Kankanhalli, M. (2024). Hallucination is Inevitable: An Innate Limitation of Large Language Models. *arXiv*. Retrieved from https://arxiv.org/abs/2401.11817.

[19] Ayala, O., & Bechard, P. (2024). Reducing hallucination in structured outputs via Retrieval-Augmented Generation. *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, 228–238. Association for Computational Linguistics. DOI: 10.18653/v1/2024.naacl-industry.19.

[20] J. Zhao et al., “LoRA Land: 310 Fine-tuned LLMs that Rival GPT-4, A Technical Report,” arXiv (Cornell University), Apr. 2024, doi: 10.48550/arxiv.2405.00732.

[21] K. P. V. Srinivasan, P. Gumpena, M. Yattapu, and V. H. Brahmbhatt, "Comparative Analysis of Different Efficient Fine Tuning Methods of Large Language Models (LLMs) in Low-Resource Setting," *arXiv*, 2024. [Online]. Available:https://arxiv.org/abs/2405.13181.

[22] G. Team et al., “GemMa: Open models based on Gemini research and technology,” arXiv (Cornell University), Mar. 2024, doi: 10.48550/arxiv.2403.08295

[23]]P. S. H. Lewis et al., “Retrieval-Augmented Generation for Knowledge-Intensive NLP tasks,” Neural Information Processing Systems, vol. 33, pp. 9459–9474, May 2020, [Online]. Available: https://proceedings.neurips.cc/paper/2020file/6b493230205f780e1bc26945df7481e5-Paper.pdf

[24] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, M. Wang, and H. Wang, "Retrieval-Augmented Generation for Large Language Models: A Survey," *arXiv preprint arXiv:2312.10997*, 2024. [Online]. Available: https://arxiv.org/abs/2312.10997

[25] H. -W. Hu, Y. -c. Lin, C. -H. Chia, E. Chuang and Y. Cheng Ru, "Leveraging Large Language Models for Generating Personalized Care Recommendations in Dementia," *2024 IEEE International Workshop on Electromagnetics: Applications and Student Innovation Competition (iWEM)*, Taoyuan County, Taiwan, 2024, pp. 1-4, doi: 10.1109/iWEM59914.2024.10649066.

[26] G. Izacard, P. Lewis, M. Lomeli, L. Hosseini, F. Petroni, T. Schick, J. Dwivedi-Yu, A. Joulin, S. Riedel, and E. Grave, "Atlas: Few-shot Learning with Retrieval Augmented Language Models," arXiv preprint, 2022. [Online]. Available: https://arxiv.org/abs/2208.03299

[27] N. Thakur, N. Reimers, A. Rücklé, A. Srivastava, and I. Gurevych, "BEIR: A Heterogenous Benchmark for Zero-shot Evaluation of Information Retrieval Models," arXiv preprint, 2021. [Online]. Available: https://arxiv.org/abs/2104.08663

[28] Y. Tang and Y. Yang, "MultiHop-RAG: Benchmarking Retrieval-Augmented Generation for Multi-Hop Queries," arXiv preprint, 2024. [Online]. Available: https://arxiv.org/abs/2401.15391

[29] Zhang, Tianjun, et al. "Raft: Adapting language model to domain specific rag." *arXiv preprint arXiv:2403.10131* (2024).