BACHELOR OF SCIENCE IN COMPUTER SCIENCE AND ENGINEERING

TRL: A Transformer Based Approach to Mental Health Counseling with Retrieval Augmented Generation and Low Rank Adaptation

Md. Zunaid Ul Alam
200041218
Shams Farhan Ivan
200041242
Zunaira Sultan
200041217

Department of Computer Science and Engineering
Islamic University of Technology
September, 2025

TRL: A Transformer Based Approach to Mental Health Counseling with Retrieval Augmented Generation and Low Rank Adaptation

Md. Zunaid Ul Alam
200041218
Shams Farhan Ivan
200041242
Zunaira Sultan
200041217

Department of Computer Science and Engineering
Islamic University of Technology
September, 2025

Declaration of Candidate

This is to certify that the work presented in this thesis is the outcome of the analysis and experiments carried out by **Md. Zunaid Ul Alam**, **Shams Farhan Ivan**, and **Zunaira Sultan** under the supervision of **Tareque Mohmud Chowdhury**, Assistant Professor, Department of Computer Science and Engineering and co-supervision of **Sabrina Islam**, Lecturer, Department of Computer Science and Engineering, Islamic University of Technology, Dhaka, Bangladesh. It is also declared that neither this thesis nor any part of it has been submitted anywhere else for any degree or diploma. Information derived from the published and unpublished work of others have been acknowledged in the text and a list of references is given.

Tareque Mohmud Chowdhury
Assistant Professor
Department of Computer Science and Engineering
Islamic University of Technology (IUT)
Date: September 29, 2025

Sabrina Islam
Lecturer
Department of Computer Science and Engineering
Islamic University of Technology (IUT)
Date: September 29, 2025

Md. Zunaid Ul Alam
Student ID: 200041218
Date: September 29, 2025

Shams Farhan Ivan
Student ID: 200041242
Date: September 29, 2025

Zunaira Sultan
Student ID: 200041217
Date: September 29, 2025

Dedicated to our honorable supervisors and faculty members, without whom this journey would have been impossible

Contents

List of Figures

List of Tables

List of Abbreviations

LLM	Large Language Model
LSTM	Long Short-Term Memory
ViT	Vision Transformers
PEFT	Parameter-Efficient Fine-Tuning
LoRA	Low-Rank Adaptation
RAG	Retrieval-Augmented Generation
ROUGE	Recall-Oriented Understudy for Gisting Evaluation
BLEU	Bilingual Evaluation Understudy

Acknowledgement

We would like to express our deepest gratitude to everyone who supported us throughout the course of this research.

First and foremost, we are profoundly grateful to our supervisors, Tareque Mohmud Chowdhury sir and Sabrina Islam ma'am, for their continuous guidance, insightful feedback, and unwavering encouragement. Their expertise and patient mentorship were instrumental in shaping the direction and quality of our work.

Our appreciation extends to the faculty and staff of the Computer Science Department for providing a supportive academic environment and access to computational resources. We are also thankful to the developers and maintainers of the open-source libraries and frameworks (including PyTorch, Hugging Face Transformers, and PEFT) that made our experiments feasible.

Finally, we acknowledge the love and encouragement of our families and friends, whose patience and understanding allowed us to focus on this work. Their steadfast support was a constant source of motivation.

Thank you all for making this journey possible.

Abstract

Mental health is a critical component of global well-being, yet access to timely and affordable care remains limited due to stigma, scarcity of professionals, and economic or geographic barriers. In Bangladesh, where prevalence rates of anxiety and depression among adolescents and university students are alarmingly high, these challenges are particularly severe. At the same time, advances in artificial intelligence (AI) and natural language processing (NLP) have created new opportunities to deliver accessible and scalable mental health support through conversational agents.

This research investigates the development of an AI-powered counseling chatbot using textual transformers. Building on the strengths of large language models, we refine response generation through the integration of Retrieval-Augmented Generation (RAG) and Low-Rank Adaptation (LoRA). RAG enables dynamic access to relevant external knowledge, while LoRA provides efficient fine-tuning, resulting in a lightweight yet effective model. Our approach emphasizes improving contextual accuracy, mitigating hallucinations, and ensuring sensitivity to the nuanced nature of mental health conversations.

The contributions of this work are threefold: (1) advancing transformer-based methods to generate precise, compassionate, and personalized counseling responses; (2) addressing key challenges such as dataset authenticity, preprocessing for sensitive text, and factual reliability; and (3) proposing a scalable framework for deploying mental health chatbots in resource-constrained contexts. Ultimately, this research aims to bridge the gap between individuals in need of psychological support and the care they are often unable to access, offering an ethical, efficient, and practical AI-driven solution.

Chapter 1

Introduction

Since mental health affects interpersonal connections, societal productivity, and personal quality of life, it has become a crucial component of global well-being. There are still major obstacles to receiving prompt, individualized mental health care, even in the face of growing public awareness and de-stigmatization initiatives. Social stigma, regional restrictions, a lack of qualified specialists, and exorbitant expenses keep many people from getting the help they require. The pervasive usage of digital technology has led to the rise of anxiety, depression, social isolation and inadequacy, specially among adolescents and young adults. Economic challenges further exacerbate psychological distress, along with workplace related stress and burnouts.

In Bangladesh, approximately 16.8% of adults aged 18-99 years and 13.6% of children aged 7-17 years suffer from mental health disorders.[13].According to a study, 33.2% of students who took university admission exams reported having anxiety, and 53.8% reported having depression. Increased mental health issues were linked to factors like living in an urban area and having a personal or family history of COVID-19 infections.[9]. Of Dhaka's 13 to 18-year-old students, 18.1% expressed moderate to severe anxiety, and 26.5% reported moderate to severe depression. Higher rates of anxiety and sadness were linked to factors including smoking and poor sleep quality.[8].

In Bangladesh, these challenges are more difficult to deal with due to the scarcity of mental health professionals and limited access to care. Moreover, persistent stigma surrounding mental health continues to deter individuals from seeking necessary support.

1.1 Introductory Information

Our research focuses on the use of textual transformers, a class of machine learning models that excel at generating human-like responses given appropriate prompts. These transformers, exemplified by models like ChatGPT, are capable of understanding and responding to text in a way that simulates natural conversation. In recent years, these models have shown significant promise in various applications, including creating chatbots capable of offering guidance and assistance in various domains.

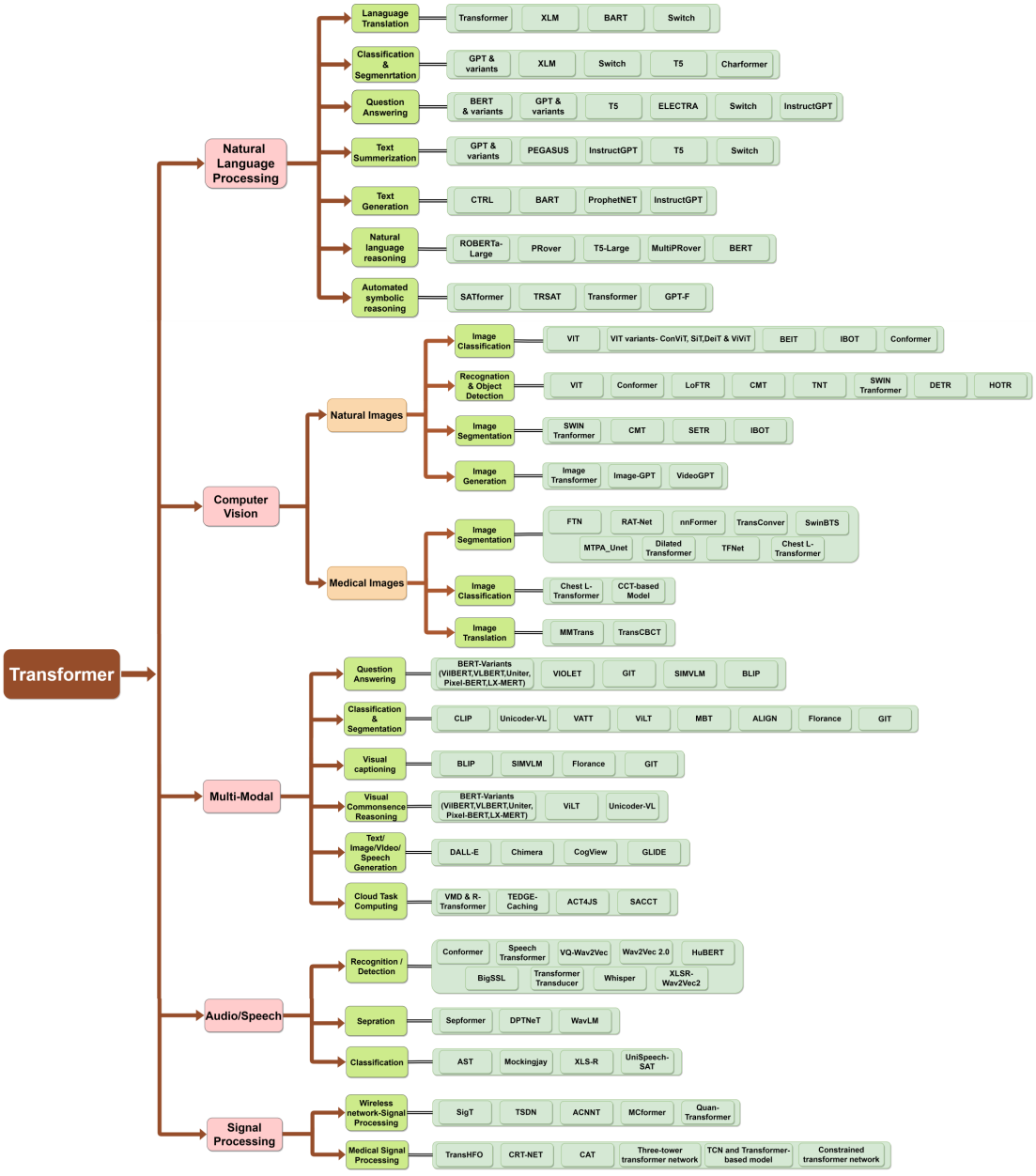


Figure 1.1: Various uses of transformers

One area where we see significant potential for improvement is in the field of mental

health care. Therapy using counsel chatbot falls under the category of Transformers -> Natural Language Processing -> Natural Language Reasoning [10]. Therapy, though incredibly valuable, can be costly and often difficult for many individuals to access. Our research seeks to address this issue by developing an advanced chatbot that provides accessible and effective counseling to people suffering from mental illness. By using recent refinement techniques in textual transformers, we aim to enhance the quality of responses generated by these models, allowing them to provide compassionate and accurate guidance to users.

This research is particularly relevant in the context of the growing role of artificial intelligence (AI) in healthcare. As AI technologies continue to evolve, they offer new possibilities for improving mental health support and expanding access to care. By combining state-of-the-art natural language processing (NLP) techniques with therapeutic models, we hope to create a system that can provide an effective alternative to traditional therapy methods, making mental health support more accessible to the general public. This work not only advances the field of AI but also has the potential to improve the well-being of individuals worldwide.

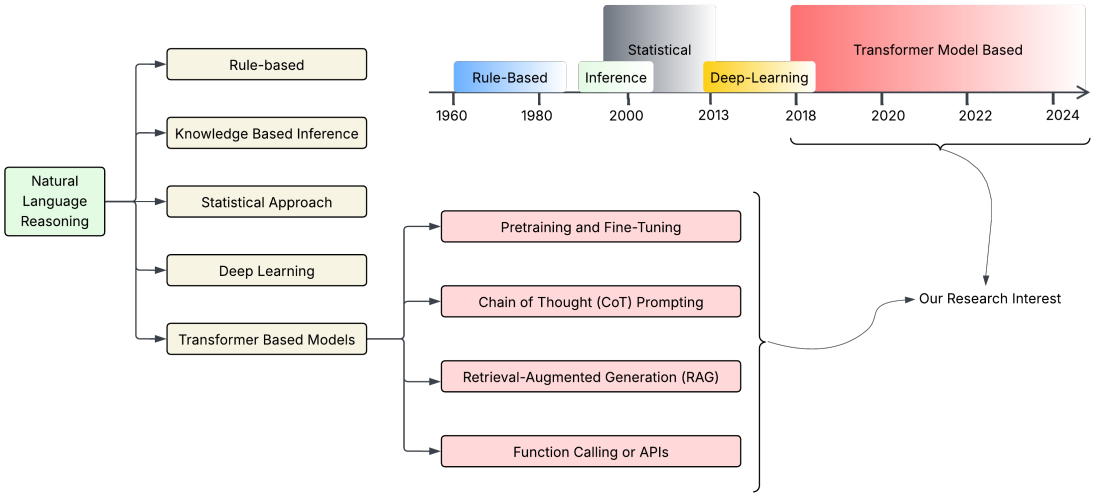


Figure 1.2: Our focus

1.2 Motivation and Scope

The motivation behind this research stems from a deeply personal and societal need: the growing challenge of accessing mental health care. Mental illness affects millions of people worldwide, and yet, despite the availability of therapies and counseling, many individuals are unable to obtain the support they need due to cost, location, or other barriers. We have witnessed firsthand how people can suffer due to the lack

of accessible mental health resources, and this has driven me to explore alternative solutions. Our motivation lies in creating a system that can provide guidance and support in a way that is both accessible and cost-effective, offering comfort to those who might otherwise go without help.

The increasing reliance on artificial intelligence (AI) and natural language processing (NLP) technologies in various sectors, including healthcare, presents an opportunity to bridge the gap between individuals in need and the support they seek. While chatbots have made significant strides in areas like customer service and personal assistants, their application in mental health care has been limited. By refining these AI models and integrating state-of-the-art techniques such as Retrieval-Augmented Generation (RAG) and Low-Rank Adaptation (LoRA), this research aims to improve the quality of responses in therapeutic contexts, making them more relevant, compassionate, and accurate.

Previous research has explored the potential of AI in mental health, with studies focusing on the use of chatbots for basic emotional support or mental health diagnostics. However, these studies often fall short in terms of delivering in-depth guidance or addressing the complexities of human emotion. Furthermore, many existing systems are limited by their inability to adapt to new data or provide responses that are personalized and contextually appropriate. This research seeks to address these shortcomings by leveraging cutting-edge AI techniques to build a more effective and adaptive counseling bot.

The scope of this research is focused on improving the performance of textual transformers in mental health applications, particularly by refining response quality and ensuring that the models remain sensitive to the nuanced nature of mental health conversations. The goal is not to replace human therapists but to offer an additional tool that can provide valuable support. By addressing gaps in existing models and applying innovative solutions, this research aims to contribute to a future where mental health care is more accessible to everyone, regardless of their circumstances.

1.3 Problem Statement

The objective of this research is to develop a transformer-based counseling assistant capable of generating responses that are accurate, empathetic, and safe in the context of mental health queries. To ensure reliability and minimize hallucinations, the model incorporates Retrieval-Augmented Generation (RAG), grounding its outputs in authoritative mental health guidelines and resources. Training efficiency is achieved

through Low-Rank Adaptation (LoRA), which enables fine-tuning with limited computational resources while maintaining strong performance. To address the issue of catastrophic forgetting and support continual adaptation, parameter-efficient fine-tuning strategies are applied, allowing the model to integrate new knowledge without compromising existing capabilities. The effectiveness of the system is evaluated through both quantitative metrics, including ROUGE, BLEU, and BERTScore, and qualitative assessments, such as LLM-as-a-judge evaluations and human surveys that measure empathy, guidance, reassurance, and appropriate self-disclosure.

1.4 Research Challenges

When applying textual transformers to mental health counseling, several research challenges arise. First, pre-processing becomes a crucial step to ensure that the sensitive and nuanced nature of mental health text data is preserved and appropriately handled. Additionally, data authenticity poses a significant concern; often, the origin of datasets is not disclosed, and in many cases, the data may even be fabricated, raising ethical and reliability issues. The scarcity of publicly available, high-quality datasets further limits the development and evaluation of models in this domain. Lastly, one of the most critical challenges is the phenomenon of hallucination, where large language models (LLMs) produce responses that sound plausible but are factually incorrect or misleading. In mental health contexts, where accuracy and trustworthiness are paramount, such hallucinations can lead to serious consequences, emphasizing the need for rigorous safeguards.

1.5 Contribution

This research aims to develop an effective counseling chatbot by fine-tuning textual transformers on a dataset comprising question-answer pairs derived from open-ended therapy sessions. Our approach goes beyond basic fine-tuning by incorporating Retrieval-Augmented Generation (RAG) and Low-Rank Adaptation (LoRA) techniques to improve model performance and responsiveness. The integration of RAG allows the model to retrieve relevant information from external sources, ensuring that the responses are contextually accurate and enriched with valuable information. Meanwhile, LoRA enhances the fine-tuning process by introducing low-rank updates, enabling the model to adapt quickly while maintaining computational efficiency. These innovations contribute to a more reliable, lightweight, and accurate model for mental health counseling, addressing the challenges of hallucination, data authenticity, and

pre-processing in this sensitive domain. Our work advances the state of the art by making it feasible to deploy a counseling bot that is both trustworthy and efficient, with a particular focus on the ethical and practical considerations inherent in mental health applications.

1.6 Roadmap

In this section, we will begin by examining the history and evolution of the technologies used to generate responses to prompts, providing a foundational understanding of how these models work. We will then trace the key advancements that have been made to enhance the basic architecture, exploring the step-by-step improvements that have shaped current approaches. Along the way, we will identify the limitations of each technology and how researchers have successfully addressed these challenges. Building on the insights from these advancements and related works, we will explore potential solutions to the ongoing issues in the field, particularly those affecting the generation of responses that function effectively as therapeutic guidance. This roadmap will guide our exploration of the state-of-the-art techniques and how they can be leveraged to improve the quality and applicability of chatbot responses in counseling contexts.

Chapter 2

Related Works

In this chapter we will discuss about the progress of current technology in the domain of mental health counselling. Many approaches and solutions have developed slowly over the time that overcomes many obstacles and issues/disadvantages that were prevalent earlier. But each progress also brought forth many other issues that have yet to be overcome.

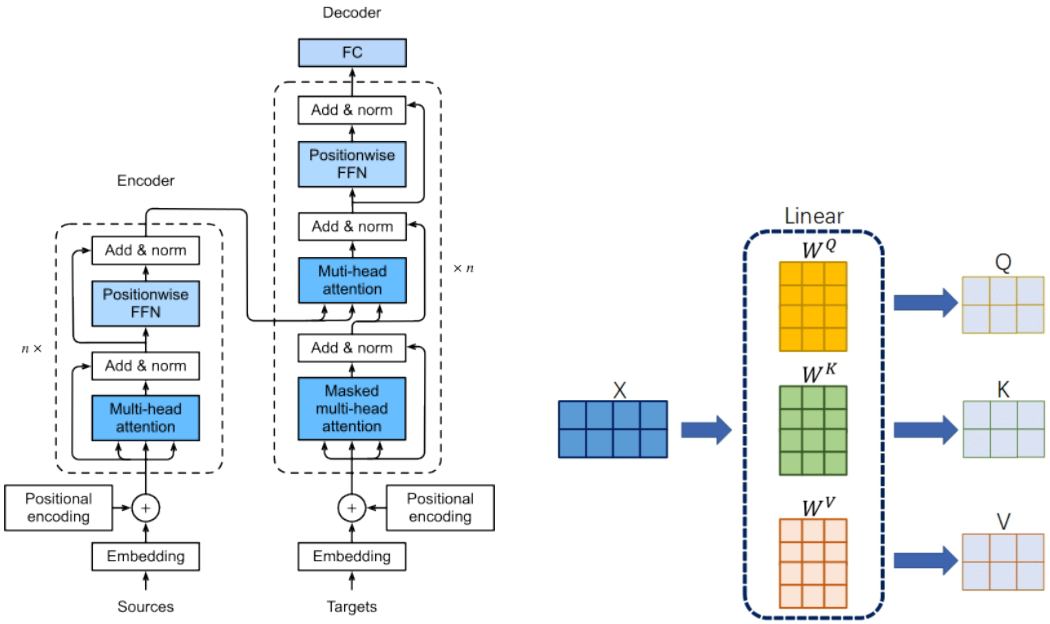
2.1 Transition from different approaches

In Figure 1.2, we observe how the concept of automated response generation has evolved: starting from rule-based systems, moving to statistical methods, and finally advancing to deep learning models, with transformer-based architectures becoming the latest and most dominant approach within the field.

Deep learning techniques initially employed Long Short-Term Memory (LSTM) models, which were capable of generating word embeddings [5]. These embeddings are vectors composed of real numbers that represent the meanings of words. However, LSTM-based architectures presented significant limitations. Firstly, learning was sequential (seq2seq), meaning that words had to be processed one after another, leading to slower training and limited parallelization. Secondly, LSTM models generated static word embeddings: each word was associated with a single, fixed vector, regardless of context. This posed challenges when dealing with polysemous words, whose meanings depend on context. For example, the word "apple" can refer both to a fruit and to a technology company. Ideally, the meaning of "apple" should vary depending on the sentence it appears in. However, the static embeddings produced by LSTMs blended all meanings into one, making it difficult for the model to distinguish between

different usages of the same word. These limitations revealed the need for models that could create **contextual** representations of words.

In 2017, a groundbreaking paper by Google researchers introduced transformer-based models [17], which addressed the key shortcomings of LSTM architectures. Transformers generate **contextual** word embeddings, meaning that the representation of a word changes dynamically based on the surrounding words in the sentence. This allows models to capture the intended meaning of words far more accurately.



(a) Original Transformer Architecture **(b)** Contextual Word Embedding Generation

Figure 2.1: Illustrations of Transformer Architecture and Contextual Embedding Generation

The core concept underlying transformers is known as *Attention*. In this mechanism, each word is associated with Key, Value, and Query vectors. The Query vector interacts with all other positions in the sequence to compute attention scores, determining how relevant each word is to the current one. The Key vector helps assess this relevance, and the Value vector contains the information used to generate the final output. The similarity between a Query and a Key determines how much attention is assigned to the corresponding Value. The higher the attention score, the more influence a particular Value vector has on the output. Through this mechanism, transformers enable efficient and context-sensitive understanding of language.

2.2 Advances in Retrieval and Parameter-Efficient Fine-Tuning

Beyond the application of transformers in mental health, recent advances in natural language processing have introduced new techniques to address critical challenges such as hallucination, factual grounding, and computational efficiency. Retrieval-Augmented Generation (RAG), introduced by Lewis et al. in 2020 **lewis2020retrieval**, combined large language models with external knowledge retrieval to generate more accurate and grounded responses. By incorporating relevant passages from external sources during the generation process, RAG reduced reliance on parametric memory alone, thereby mitigating hallucinations and enhancing faithfulness in generated outputs. Since its introduction, RAG has been widely adopted in open-domain question answering, dialogue systems, and specialized applications where factual consistency is essential.

Complementing this, Hu et al. in 2021 [6] proposed Low-Rank Adaptation (LoRA), a parameter-efficient fine-tuning method designed to address the high computational costs associated with updating large-scale models. Instead of retraining all model parameters, LoRA injects trainable low-rank decomposition matrices into existing weights, significantly reducing resource requirements while maintaining performance. This approach has proven especially effective in low-resource settings and domain-specific tasks, where full fine-tuning is impractical.

Despite their success in broader NLP domains, the integration of RAG and LoRA in the mental health sector remains underexplored. While prior works have focused on improving empathy, domain adaptation, or data augmentation, few studies have directly addressed the combined challenge of ensuring factual reliability and computational efficiency in therapeutic contexts. This research aims to bridge that gap by leveraging RAG for knowledge-grounded responses and LoRA for efficient fine-tuning, creating a more adaptable and trustworthy counseling chatbot.

2.3 Transformer Use in the Mental Health Sector

Recent studies have leveraged transformer-based architectures to address mental health challenges across various domains. In 2022, Trappey et al. [16] introduced an empathy-centric counseling chatbot system capable of sentimental dialogue analysis, aiming to provide psychological support through empathetic conversations. Although the system effectively reduced psychological sensitivity and stress among users, its limitation

lies in the lack of significant improvement in life-impact measures, indicating a gap in addressing deeper behavioral and emotional shifts.

In 2024, Adhikary et al. [1] explored the ability of large language models to summarize mental health counseling sessions, proposing benchmark datasets and evaluation strategies. Their work demonstrated the potential for automating counseling session summaries, yet it faced challenges concerning faithfulness, hallucination, and limited domain-specific tuning, which could affect the reliability of generated summaries in sensitive clinical contexts.

Guo et al. [4] provided a systematic review in 2024 of large language models for mental health applications, highlighting their advantages in empathetic response generation, early diagnosis, and therapeutic support. Nonetheless, major concerns were identified regarding model bias, privacy breaches, and lack of real-world deployment validation.

Building upon prediction tasks, Xu et al. [18] presented Mental-LLM in 2024, an approach leveraging large language models for mental health prediction via online text data. Although the model showcased strong predictive performance, it encountered limitations in generalizability across different demographics and linguistic styles, exposing issues of fairness and inclusivity.

In 2023, Liu et al. [12] proposed ChatCounselor, which fine-tuned large language models to provide mental health support conversations. Despite advancements in generating supportive responses, challenges remained around factual consistency, therapeutic accuracy, and model interpretability, especially when deployed in high-stakes environments.

Data augmentation for counseling conversations was addressed by Kim et al. [11] in 2024, who proposed a pipeline utilizing large language models to enhance low-resource datasets. While effective in expanding training data, the work encountered difficulties ensuring that synthetically generated dialogues preserved therapeutic realism and linguistic diversity.

Privacy and domain specialization were the focus of Zhang and Luo [19] in 2024, who integrated dual-memory systems and privacy safeguards within conversational psychotherapy models. Despite offering innovative mechanisms, their system demanded significant computational resources and faced obstacles in achieving seamless memory retention and adaptation to dynamic conversational contexts.

Other notable efforts include the 2024 study by Pant and Mukhiya [14], where the

reflective listening capabilities of GPT-4 and GPT-4-Turbo were evaluated using the CounselChat dataset. Although the models demonstrated promising results in empathy simulation, inconsistencies in nuanced reflective response generation highlighted ongoing challenges.

Additionally, Patil and Rasave [15] introduced an AI chatbot for counseling therapy in 2021, focusing on basic therapeutic support. However, the chatbot's rule-based nature limited its adaptability and depth of empathetic engagement compared to transformer-based counterparts.

Corpus creation efforts such as Iqbal et al. [7] in 2022 and Banshal et al. [3] in 2023 developed emotion-labeled datasets in regional languages like Bengali, which are essential for low-resource settings. Nevertheless, the corpora face scalability limitations and challenges associated with annotation subjectivity.

Earlier foundational work by Althoff et al. [2] in 2016 applied natural language processing techniques to counseling conversation datasets, offering large-scale empirical insights into counselor-client interactions. While impactful, these early approaches lacked the sophistication of transformer architectures for capturing contextual nuances.

Several overarching surveys, including Islam et al. [10] in 2023 and Vaswani et al. [17] in 2023, underscored the transformative role of attention mechanisms and transformers in deep learning tasks. However, they acknowledged persistent issues related to model explainability, computational demand, and performance on small-scale, specialized datasets.

Despite the significant progress outlined above, existing approaches frequently encounter limitations related to hallucination, insufficient domain adaptation, privacy concerns, and difficulties handling low-resource or specialized mental health data. To address these challenges, our proposed model introduces retrieval-augmented generation (RAG) to ground model outputs on relevant external knowledge sources, thereby enhancing faithfulness and reducing hallucination. Furthermore, we apply Low-Rank Adaptation (LoRA) techniques to enable efficient domain-specific fine-tuning, improving adaptability to specialized therapeutic contexts while mitigating the high computational costs associated with full model retraining.

Chapter 3

Proposed Methodology

3.1 System Pipeline

The proposed counseling assistant follows a structured pipeline designed to generate accurate, empathetic, and safe responses. The process begins when a user submits a counseling-related query, which is encoded into dense vector representations using an embedding model. At the same time, external documents containing authoritative mental health guidelines (e.g., WHO, IASC, mhGAP) are embedded into the same vector space.

A Retrieval-Augmented Generation (RAG) module constructs a vector database of these external knowledge sources and retrieves the top-k most relevant passages. These passages are concatenated with the original user query to provide grounded context. The enriched query is then passed to a base large language model (Vicuna, LLaMA, Mistral, or Falcon), where the decoder integrates both the retrieved evidence and the user's input to generate a response.

To enable resource-efficient fine-tuning, Low-Rank Adaptation (LoRA) modules are incorporated. LoRA inserts low-rank adapters into the model, enabling parameter-efficient updates while mitigating catastrophic forgetting. This makes the system lightweight, adaptable to new data, and feasible to train with limited resources. The final output is a balanced counseling response, simultaneously informative and empathetic.

3.2 Training Strategies

In addition to architectural design, experiments were conducted with different training strategies on the CounselChat dataset. An initial attempt involved weight-based training, where the number of upvotes for a particular answer served as an importance weight.

3.2.1 Oversampling

The first method applied oversampling by duplicating samples in proportion to their upvotes. However, this introduced severe data imbalance. The transformer model memorized duplicated samples, which reduced its generalization ability and caused frequent hallucinations in generated outputs.

3.2.2 Weighted Loss Function

To overcome these issues, a weighted loss formulation was adopted, inspired by multi-task learning objectives used in object detection and classification. After normalizing upvote counts, weights were applied to each training instance using a logarithmic scaling function. The modified training objective was defined as follows:

$$\mathcal{L}_{base} = \text{criterion}(outputs, target) \tag{3.1}$$

$$w_i = \log(1 + u_i) \tag{3.2}$$

$$\mathcal{L}_{weighted} = \frac{\sum_{i=1}^{N} \mathcal{L}_{base}^{(i)} \cdot w_i}{\sum_{i=1}^{N} w_i + 10^{-8}} \tag{3.3}$$

where u_i represents the upvote count for sample i, and $\mathcal{L}_{base}$ is the unweighted loss.

Sanity checks using ROUGE on highly upvoted samples demonstrated measurable improvements over the baseline, indicating that upvote-aware training helped prioritize higher-quality counseling responses.

3.3 Quantitative Evaluation Metrics

To evaluate linguistic quality, the model was tested using established natural language generation metrics.

3.3.1 ROUGE

ROUGE measures overlap between system outputs and reference texts, focusing on recall-oriented evaluation. For ROUGE-n, the recall is defined as:

$$\text{ROUGE-}n = \frac{\sum_{\text{gram}_n \in Ref} \min\left(Count_{sys}(\text{gram}_n), Count_{ref}(\text{gram}_n)\right)}{\sum_{\text{gram}_n \in Ref} Count_{ref}(\text{gram}_n)} \tag{3.4}$$

ROUGE-L additionally captures the longest common subsequence (LCS) between candidate and reference responses.

3.3.2 BLEU

BLEU evaluates precision of n-gram matches while penalizing excessively short outputs through a brevity penalty:

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{3.5}$$

where p_n is the modified n-gram precision and BP is the brevity penalty.

3.3.3 BERTScore

BERTScore uses contextual embeddings from pre-trained language models (e.g., BERT) to compute semantic similarity. Candidate and reference tokens are embedded, and similarity is measured using cosine similarity:

$$\text{BERTScore} = \frac{1}{|X|} \sum_{x \in X} \max_{y \in Y} \cos(\mathbf{e}_x, \mathbf{e}_y) \tag{3.6}$$

where $\mathbf{e}_x$ and $\mathbf{e}_y$ are contextual embeddings of candidate and reference tokens.

Initially, single-reference evaluation produced misleading results because counseling questions often have multiple valid answers. Therefore, multi-reference evaluation was adopted to provide more reliable measurements.

3.4 Qualitative Evaluation

While automatic metrics measure linguistic quality, they fail to capture empathy, reassurance, or therapeutic effectiveness. To complement quantitative evaluation, qualitative assessments were performed through LLM-as-a-judge and human-like scoring frameworks.

Large language models such as DeepSeek-R1 and Zephyr-7B were prompted to evaluate counseling responses on a five-point Likert scale across five categories: empathy, information, direct guidance, reassurance, and self-disclosure. Each score was accompanied by a short justification, followed by an overall impression.

For example, given the user message:

> *"I'm so stressed about this upcoming presentation. I feel like I'm going to fail."*

and the assistant response:

> *"It's completely normal to feel stressed before a big presentation. Try breaking your preparation into small, manageable steps, like outlining your key points first. You've prepared for this, and you know your material."*

the evaluation framework yielded the following sample scores: Empathy = 4 (validates feelings as normal), Direct Guidance = 5 (clear, actionable strategy), Reassurance = 3 (offers comfort but could be stronger). This framework ensures that both linguistic correctness and therapeutic quality are systematically assessed.

3.5 Summary

This methodology integrates retrieval-augmented generation, parameter-efficient fine-tuning, upvote-aware training, and a robust evaluation framework. By combining quantitative metrics with qualitative human-aligned evaluations, the system aims not only to produce accurate responses but also to ensure empathy, reassurance, and trustworthiness in mental health counseling scenarios.

Chapter 4

Results and Discussion

4.1 Datasets and Experimental Setup

In this research, two publicly available datasets related to psychological counseling and emotional well-being were utilized: the **Counsel-Chat** dataset and the **Psych8k** dataset.

4.1.1 Counsel-Chat Dataset

The Counsel-Chat dataset, curated by *nbertagnolli*, contains a total of 2775 question-answer pairs extracted from an online counseling forum. The dataset is entirely text-based and encompasses user-submitted questions and therapist-provided responses. These samples are categorized into 31 distinct topics, including:

- Depression (317 samples)
- Anxiety (256 samples)
- Relationships (245 samples)
- Self-esteem (198 samples)
- Trauma (176 samples)
- Grief (154 samples)
- Parenting (132 samples)
- Addiction (118 samples)
- Career Counseling (105 samples)

- Military Issues (3 samples)

The dataset shows a noticeable class imbalance, with some categories (e.g., Depression, Anxiety) being significantly more represented than others (e.g., Military Issues). To address these imbalances, stratified sampling techniques were considered during data partitioning to ensure fair evaluation across categories.
This dataset was used for weighted learning. The upvotes on each answer was used as weight to train the model for better answers

Sample Entry

Table 4.1: Example from the Counsel-Chat dataset

Question ID	2
Question Title	Do I have too many issues for counseling?
Question Text	My mother is combative with me when I say I don't want to talk with her about my depression (...)
Question Link	`https://counselchat.com/questions/i-feel-like-my-mother-doesn-t-support-me`
Topic	Depression
Therapist Info	Dr. Meredyth Lawrynce, Serving Clients Nationwide
Therapist URL	`https://counselchat.com/therapists/dr-meredyth-lawrynce`
Answer Text	Do you live with your mom and have constant interaction with her?...
Upvotes	2
Views	187

4.1.2 Psych8k Dataset

The Psych8k dataset, developed by *EmoCareAI*, consists of 8000 text entries derived from 260 real-life counseling sessions. Each entry captures short dialogues or counseling exchanges involving a wide range of emotional themes, such as:

- Family
- Relationships
- Career Development
- Academic Stress

The Psych8k dataset is also purely text-based and reflects a richer conversational context compared to Counsel-Chat. Given its origin from real counseling sessions, the data exhibits variability in language style, emotional tone, and conversation length.

No significant missing data issues were identified; however, slight topic imbalances were observed.

Sample Entry

Table 4.2: Example from the Psych8k dataset

Input	Lately, I've been feeling a bit off. I sometimes find it hard to focus and concentrate on tasks (...)
Output	I appreciate you sharing your concerns with me. It's good to hear that you can still function normally (...)
Instructions	If you are a counsellor, please answer the questions based on the description of the patient.

4.1.3 Experimental Setup

The experiments were conducted to fine-tune a causal language model using parameter-efficient tuning techniques. The computational environment consisted of an NVIDIA RTX 3090 GPU with 24GB memory. The primary software tools used were Python 3.10, PyTorch 2.1, Hugging Face Transformers 4.39, and the Hugging Face PEFT (Parameter-Efficient Fine-Tuning) library. Training was performed on Google Colab to leverage access to high-memory GPU instances.

Software Libraries

The implementation was carried out in Python, leveraging several state-of-the-art deep learning and natural language processing libraries:

- **PyTorch** (`torch`) for model implementation and tensor operations.
- **Transformers** (`transformers`) from Hugging Face for pre-trained large language models and training utilities.
- **Datasets** (`datasets`) for efficient data preprocessing and management.
- **Accelerate** (`accelerate`) for optimized training across hardware configurations.

Hyperparameters

The model was trained using a set of carefully chosen hyperparameters, informed by standard practices in fine-tuning large language models:

Table 4.3: Summary of training hyperparameters.

Hyperparameter	**Value**
Learning Rate	5×10^{-5}
Number of Epochs	3
Per-device Training Batch Size	8
Per-device Evaluation Batch Size	8
Warm-up Steps	500
Weight Decay	0.01

These hyperparameters were selected to balance efficiency and performance. The learning rate follows common recommendations for transformer fine-tuning, while the batch size was constrained by GPU memory considerations. The number of epochs was empirically determined to provide convergence without overfitting. Warm-up steps and weight decay were included to stabilize training and improve generalization, consistent with best practices reported in prior studies.

4.2 Results

4.2.1 Quantitative Results

The performance of the four base models—Vicuna-7B, LLaMA-2-7B, Mistral-7B, and Falcon-7B—was evaluated under three different conditions: zero-shot, few-shot, and fine-tuned . The results are summarized in Tables 4.4, 4.5 and 4.6.

Table 4.4: Performance Metrics for Zero Shot Training

Model	ROUGE-1	ROUGE-2	ROUGE-L	BLEU	BERT
Vicuna-7B-v1.5	0.1492	0.021	0.0715	0.0026	0.5358
Mistral-7B	0.2044	0.123	0.0825	0.0041	0.5567
LLaMA-2-7B	0.1190	0.059	0.0532	0.0025	0.4608
Falcon-7B	0.1338	0.189	0.0693	0.0036	0.4882

Overall, Mistral-7B achieved the highest quantitative scores (especially in ROUGE-1 and ROUGE-L), while LLaMA-2-7B obtained the strongest BLEU score. Vicuna-7B showed more balanced results across all metrics.

Table 4.5: Performance Metrics for Few Shot Training

Model	ROUGE-1	ROUGE-2	ROUGE-L	BLEU	BERT
Vicuna-7B-v1.5	0.2598	0.076	0.1248	0.0088	0.6165
Mistral-7B	0.2984	0.421	0.1350	0.0162	0.6571
LLaMA-2-7B	0.1913	0.072	0.1083	0.0073	0.5419
Falcon-7B	0.2210	0.321	0.1134	0.0092	0.5617

Table 4.6: Performance Metrics for Fine Tuned Training

Model	ROUGE-1	ROUGE-2	ROUGE-L	BLEU	BERT
Vicuna-7B-v1.5	0.3143	0.148	0.1643	0.0149	0.7011
Mistral-7B	0.3540	0.350	0.1817	0.0292	0.7317
LLaMA-2-7B	0.2506	0.183	0.1496	0.0154	0.6282
Falcon-7B	0.2856	0.693	0.1693	0.0175	0.6547

4.2.2 Qualitative Results

While quantitative metrics capture lexical overlap, they remain limited in evaluating the quality of counseling responses. To address this gap, we conducted qualitative evaluations using the LLM-as-a-Judge framework. Models were assessed across five dimensions: *Empathy*, *Information*, *Direct Guidance*, *Reassurance*, and *Self-Disclosure*. Figure 4.1 present the comparative results for Fine-Tuned models, Base-Model + RAG, and Fine-Tuned + RAG settings.

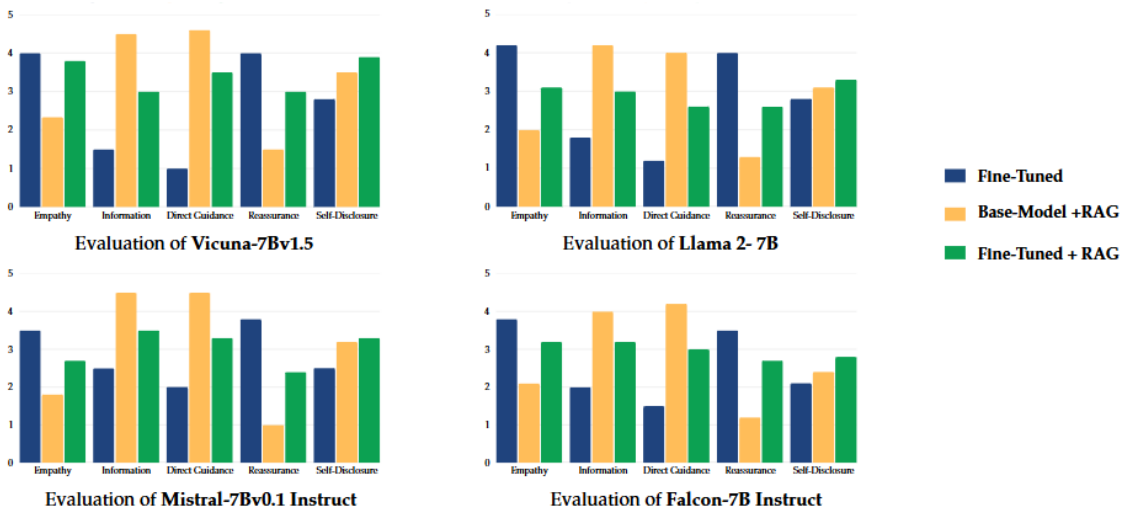


Figure 4.1: Image demonstrating qualitative analysis.

The results reveal a consistent trade-off between the approaches. Fine-tuning alone emphasized empathetic and reassuring responses, producing outputs with higher warmth but limited actionable detail. In contrast, the Base-Model + RAG configuration yielded stronger performance in terms of informativeness and direct guidance, but at the

expense of emotional resonance. The hybrid Fine-Tuned + RAG approach demonstrated the most balanced outcomes, achieving competitive scores across all five evaluation criteria. This suggests that combining fine-tuning with retrieval best supports the dual goals of counseling: being both emotionally supportive and substantively informative.

In addition to within-model comparisons, we benchmarked our approach against existing works, including Vicuna-7B, LLaMA-2, and GPT-3.5-turbo (Figure 4.2). The radar plot illustrates that the Fine-Tuned Vicuna-7B with RAG achieves the most balanced performance across all dimensions. While Vicuna-7B alone demonstrates strong empathy and reassurance, it lags in providing concrete information and direct guidance. LLaMA-2 offers more structured information but falls short in self-disclosure and reassurance. GPT-3.5-turbo provides relatively stable performance across criteria, though without excelling in any particular dimension. By contrast, the Fine-Tuned Vicuna-7B with RAG not only preserves empathetic and supportive qualities but also improves significantly in informativeness and direct guidance, underscoring its suitability for counseling contexts where both warmth and actionable content are essential.

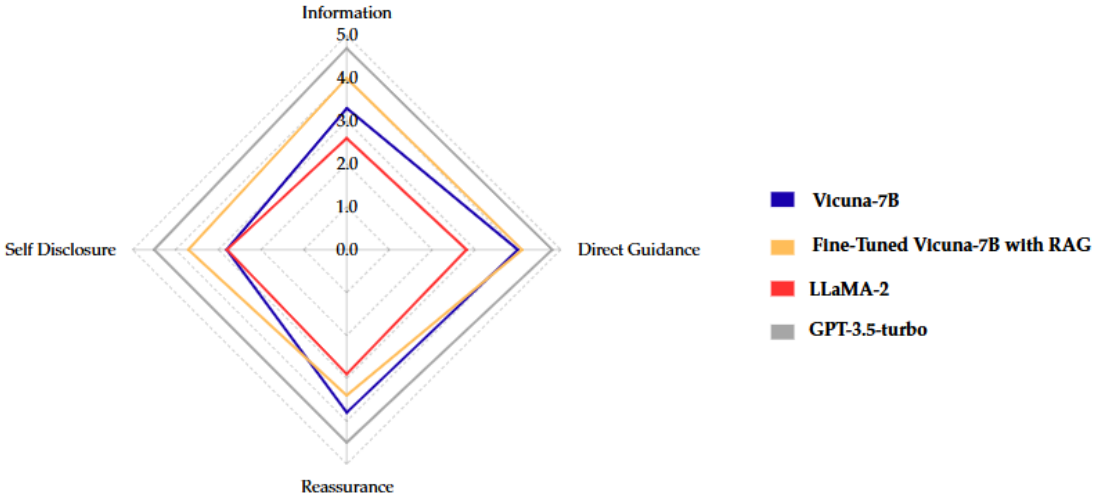


Figure 4.2: Image demonstrating Comparison of Existing Work

We further evaluated the fully trained models (Vicuna, LLaMA, Mistral, and Falcon) on the five qualitative dimensions, as shown in Figure 4.3. The results highlight notable differences in model strengths. Vicuna demonstrates the highest scores in *Empathy*, *Direct Guidance*, and *Self-Disclosure*, indicating its capacity to generate supportive and relatable responses. Mistral achieves competitive performance in *Information* and *Direct Guidance*, reflecting its strength in producing structured and informative outputs. LLaMA performs consistently across most dimensions but does not domi-

nate in any single category, suggesting a balanced yet less specialized profile. Falcon shows moderate performance overall, with its strongest outcomes in *Empathy* and *Information*, but weaker results in *Reassurance* and *Self-Disclosure*. Taken together, these results suggest that Vicuna provides the most human-like counseling qualities, while Mistral offers a strong alternative for information-centric tasks.

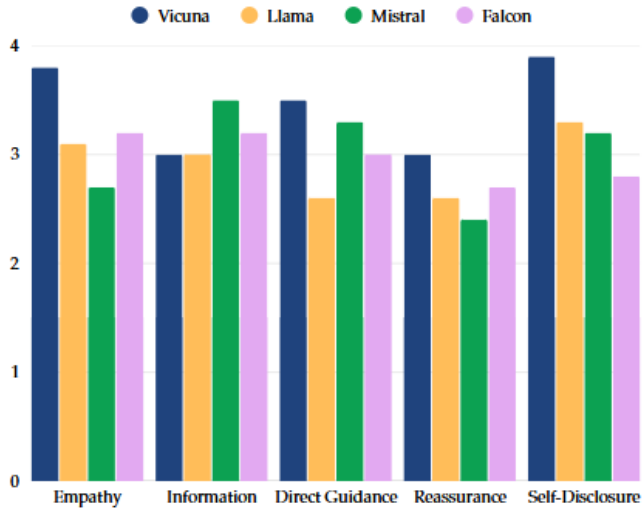


Figure 4.3: Evaluation of Fully Trained Models in Specific Criteria

To obtain a holistic view of model performance, we also computed a multiplicative score that aggregates results across all five qualitative criteria (Figure 4.4). This metric

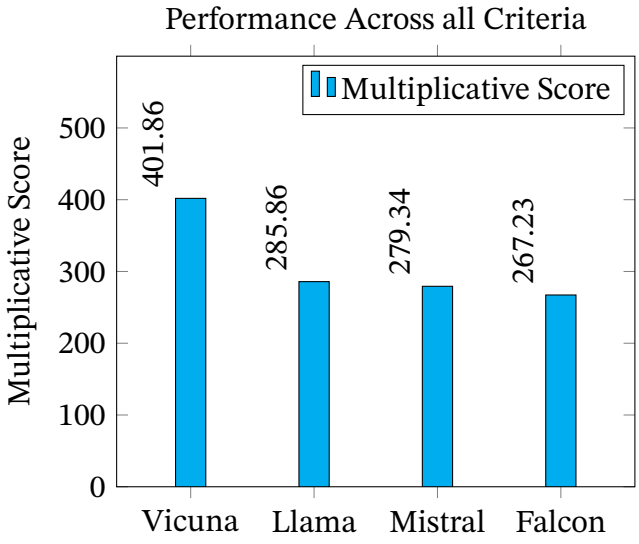


Figure 4.4: Performance Across all Criteria

rewards models that achieve consistently high performance across dimensions rather than excelling in only a subset. Vicuna achieves the highest score (401.86), reflecting its strong balance between empathy, reassurance, and self-disclosure, while still maintaining competitive levels of information and direct guidance. LLaMA (285.86) and

Mistral (279.34) follow, showing more moderate but relatively stable outcomes. Falcon records the lowest aggregate score (267.23), aligning with its weaker performance in reassurance and self-disclosure. Overall, these results reaffirm Vicuna's relative strength as a counseling model, while also highlighting the potential of LLaMA and Mistral as balanced alternatives.

Human Evaluation

To further validate results, a human evaluation was conducted using 54 participants (teachers, students, and working adults). The Fine-Tuned+RAG Vicuna-7B model achieved a win rate of 56.83% against the baseline Fine-Tuned model (43.17%).

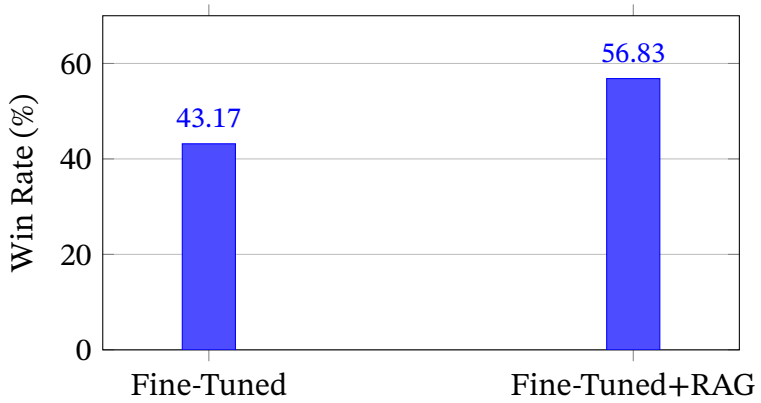


Figure 4.5: Human evaluation win rate: Fine-Tuned vs Fine-Tuned+RAG Vicuna-7B.

This finding reinforces that while automated metrics capture certain aspects of performance, human preference aligns strongly with balanced models that integrate both empathy and guidance.

4.2.3 Summary of Results

The quantitative results showed that Mistral-7B excelled in ROUGE and BLEU metrics, while Vicuna-7B demonstrated superior performance in qualitative dimensions such as empathy, reassurance, and direct guidance. This indicates a trade-off: models optimized for surface-level accuracy may lose depth in human-like qualities.

Qualitative evaluations confirmed that integrating RAG improved informativeness and guidance but slightly reduced warmth and empathy. The weighted training approach improved response quality, with human evaluators preferring Weighted+RAG outputs 56.83% of the time.

4.3 Interpreting the Results

Since our evaluation was conducted using only 100 samples, the model achieved notably higher ROUGE and BLEU scores compared to the GPT-4 and GPT-4 Turbo models reported in [14]. However, it is important to note that a direct comparison is not fully appropriate due to the limited evaluation size and differing experimental setups. What can be concluded confidently is that the fine-tuned model demonstrates significant improvement over the base (untrained) model, indicating the effectiveness of the training process.

To obtain a more definitive comparison, a comprehensive evaluation strategy would be required. Specifically, the model should be trained 5 times, each time using 20% of the full dataset, and the results should be aggregated across these runs. Until such an evaluation is performed, the final performance metrics remain undetermined.

An important observation from the results is that the *LoRA+RAG* model consistently achieves more balanced scores across all evaluation criteria. Unlike the base model, which performs extremely well in some areas (e.g., Empathy, Approval/Reassurance) but poorly in others (e.g., Information, Direct Guidance), the current model maintains steady, average-level performance across all dimensions. This indicates that the *LoRA+RAG* model is better at providing well-rounded responses, ensuring that no single aspect of counseling quality is disproportionately emphasized while others are neglected.

4.4 Challenges and Limitations

During the course of this research, several challenges and limitations were encountered that shaped both the methodology and the outcomes.
One of the primary challenges was related to handling class imbalance in the dataset. Initial attempts at oversampling combined with weighted training did not yield improvements; in fact, they led to unstable results. When we introduced weighted loss as a corrective measure, the expectation was that ROUGE and BLEU scores would marginally decrease, reflecting a trade-off in favor of more balanced learning. However, verifying this effect proved difficult. A natural way to sanity-check would have been to compare ROUGE and BLEU scores on subsets of samples with higher human vote counts, where improvements should have been more apparent. Unfortunately, a lack of established techniques for weighted evaluation limited our ability to fully validate this approach.

Another significant limitation stems from the conversational nature of therapeutic dialogue. While our setup was designed around single-prompt evaluations, real therapy sessions are inherently multi-turn, involving back-and-forth exchanges that build context and rapport over time. The single-prompt design, although practical for controlled experimentation, restricts the ecological validity of our findings by not capturing the iterative, dynamic nature of real-world mental health support.
These limitations highlight important directions for future research. Developing more reliable evaluation methods that account for weighted responses, as well as extending the system to support multi-turn conversational contexts, would strengthen both the validity and applicability of the framework. By acknowledging these challenges, we aim to provide context for our results and pave the way for more robust advancements in this field.

4.5 Conclusion of Results

This chapter presented a comprehensive evaluation of both the fine-tuned and baseline models using the Counsel-Chat and Psych8k datasets. Through quantitative metrics (ROUGE-1, ROUGE-2, ROUGE-L, BLEU) and visual comparisons, we demonstrated that the fine-tuning process led to consistent improvements over the untrained model across all evaluation measures. Although the evaluation was conducted on a limited sample size of 100 samples, the results are promising and indicate that the fine-tuned model better captures the nuances of counseling dialogue compared to its base counterpart.

When compared to the GPT-4 and GPT-4 Turbo models reported in [14], our model achieved notably higher BLEU scores and comparable ROUGE scores. However, due to differences in evaluation procedures and dataset sizes, a definitive comparison cannot yet be drawn. A more rigorous validation approach—training the model multiple times on different dataset splits and aggregating the results—would be necessary for a complete assessment of the model's final capabilities.

Beyond baseline comparisons, our experiments with multiple LLMs revealed distinct strengths and trade-offs. **Mistral-7B** achieved the best quantitative performance in terms of ROUGE and BLEU scores, highlighting its ability to align closely with reference responses. **Vicuna-7B**, on the other hand, consistently outperformed others in qualitative dimensions such as empathy, reassurance, and self-disclosure, making it particularly well-suited for emotionally sensitive counseling tasks. **LLaMA-2-7B** performed competitively, achieving the highest BLEU score among the models, while

Falcon-7B delivered moderate results across both quantitative and qualitative metrics. These comparisons underscore that the optimal choice of model depends on whether the priority is factual informativeness (Mistral) or empathetic, supportive communication (Vicuna).

A key finding across all models is the observed trade-off between empathy and informativeness. Since the datasets were primarily assurance-oriented, LoRA fine-tuning excelled at producing empathetic and supportive responses. However, these often lacked structured guidance. Integrating Retrieval-Augmented Generation (RAG) enriched the outputs with factually accurate and detailed content, but sometimes reduced the warmth of responses. The combined RAG–LoRA framework successfully balanced these two aspects, producing answers that were both empathetic and informative. This hypothesis was tested and confirmed through the evaluation, highlighting the nuanced interplay between dataset characteristics, fine-tuning strategies, and response quality.

Importantly, the adoption of lightweight fine-tuning techniques such as LoRA substantially reduced computational costs compared to training full-scale models like GPT-4. Although further experimentation is needed to refine the trade-off between empathy and informativeness, these results suggest that cost-effective domain-specific adaptation is achievable, particularly when combined with knowledge-grounding through RAG.

Looking ahead, future work should focus on enhancing personalization in responses. Therapists typically adapt their communication based on the client's traits and emotional state. To approximate this, conversational agents could integrate personality-aware modeling, where a lightweight personality assessment precedes the user's prompt to guide response generation. This would likely require multiple specialized models working in tandem to produce responses that are both emotionally supportive and contextually precise.

Overall, these findings underscore the value of targeted fine-tuning for specialized tasks while balancing performance gains with computational efficiency. By comparing across multiple models and analyzing their respective strengths, this study highlights that no single model is universally superior. Instead, hybrid approaches such as LoRA+RAG, coupled with model selection based on task priorities, provide the most promising path toward safe, empathetic, and effective AI counseling systems.

Chapter 5

Conclusion

The primary objective of this research was to investigate whether fine-tuning a lightweight transformer-based model could meaningfully improve counseling response generation while maintaining significantly lower computational costs compared to large-scale models such as GPT-4. To achieve this, we explored different training strategies, including LoRA fine-tuning, Retrieval-Augmented Generation (RAG), and their hybrid combination, and compared them against both untrained baselines and state-of-the-art large models.

5.1 Summary of Findings

The results of this study provide strong evidence in support of lightweight fine-tuning for specialized counseling applications:

- **Improvement over the untrained baseline:** The fine-tuned model consistently outperformed the untrained model across all evaluation metrics (ROUGE-1, ROUGE-2, ROUGE-L, BLEU), confirming the effectiveness of task-specific fine-tuning for enhancing response quality.
- **Comparison with large models:** Despite the significant difference in scale, our fine-tuned model achieved higher BLEU scores and comparable ROUGE scores compared to GPT-4 and GPT-4-Turbo [14]. This demonstrates that targeted fine-tuning can yield competitive performance relative to state-of-the-art models, while requiring a fraction of the computational resources.
- **LoRA vs. RAG vs. Hybrid:** LoRA fine-tuning excelled in empathy and reassurance, whereas RAG performed better in providing information and direct

guidance. The hybrid LoRA+RAG model struck a balance, achieving more stable scores across all dimensions. This balanced performance was further supported by comparative win-rate and higher-score frequency analyses, where the hybrid model consistently outperformed the LoRA-only model.

- **Computational efficiency:** Lightweight fine-tuning with LoRA proved to be highly resource-efficient compared to training large models, making it a cost-effective and sustainable approach for developing domain-specific AI.
- **Trade-offs and limitations:** A recurring theme in the results was the trade-off between empathy and informativeness. While LoRA prioritized emotionally supportive responses, RAG improved factual grounding. The hybrid model mitigated this trade-off, but achieving the optimal balance remains an open challenge.

5.2 Contributions

This work makes the following contributions:

- A comprehensive evaluation of lightweight fine-tuning for counseling dialogue generation on two real-world datasets (CounselChat and Psych8k).
- An empirical comparison of untrained, fine-tuned, LoRA-only, RAG-only, hybrid LoRA+RAG models, and large-scale GPT-4 variants.
- Insights into the trade-offs between empathy, reassurance, guidance, and informativeness in counseling responses, supported by both quantitative metrics and qualitative, LLM-as-a-judge evaluations.
- A demonstration of how lightweight fine-tuning can drastically reduce computational costs while retaining competitive performance.

5.3 Implications

The broader implication of this work is that efficient and accessible fine-tuning methods can democratize the development of specialized AI systems in domains like mental health support, where deploying large-scale models may be impractical. By enabling cost-effective, domain-adapted models, this research paves the way for more equitable and responsible use of AI in sensitive,

5.4 Future Work

While this study demonstrates the potential of lightweight fine-tuning for counseling dialogue generation, several avenues remain open for further exploration and validation:

- **Comprehensive evaluation:** A complete 5-fold validation strategy, where the model is repeatedly trained on different 20% splits of the dataset and results are aggregated, is needed to establish robust and definitive performance metrics.
- **Human evaluation with domain experts:** Although LLM-as-a-judge provided scalable assessments, future work must include human evaluation by trained counselors or psychologists to ensure clinical reliability and practical applicability.
- **Multi-turn dialogue system:** Real-world counseling is inherently interactive and contextual. Future extensions should develop multi-turn dialogue systems with context memory and history tracking to better capture the dynamics of therapeutic conversations.
- **Personalization:** Counselors naturally adapt their style to individual clients. Future research should explore personality-aware modeling, where an initial assessment (e.g., via a lightweight personality test) informs and personalizes response style throughout the interaction.
- **Enhancing retrieval:** To broaden informativeness, the RAG system can be expanded to include more diverse and domain-specific documents. Future work should also compare different retrieval strategies (e.g., dense retrieval, hybrid retrieval) to optimize knowledge grounding.
- **Balancing empathy and informativeness:** A key challenge identified in this study is the trade-off between empathetic and informative responses. Future systems could dynamically balance these dimensions using adaptive weighting schemes or reinforcement learning with human feedback.
- **Weighted evaluation using LLM scoring:** Instead of treating all samples equally, future evaluations could incorporate LLM-based quality scoring as weights, giving more importance to responses judged as high-quality by domain experts or language models.
- **Exploration of lightweight adaptation techniques:** Beyond LoRA, additional efficient fine-tuning strategies such as adapters, prefix-tuning, and quantization-

aware methods could be investigated to further optimize performance while keeping computational costs low.

In summary, future work should focus on strengthening evaluation through expert validation, expanding retrieval and personalization, and extending the system toward interactive, multi-turn dialogue. These directions will help create more robust, context-aware, and clinically meaningful AI counseling assistants.

References

[1] P. K. Adhikary et al., "Exploring the efficacy of large language models in summarizing mental health counseling sessions: Benchmark study," *JMIR Mental Health*, vol. 11, e57306, 2024.

[2] T. Althoff, K. Clark, and J. Leskovec, "Large-scale analysis of counseling conversations: An application of natural language processing to mental health," *Transactions of the Association for Computational Linguistics*, vol. 4, pp. 463–476, 2016.

[3] S. K. Banshal, S. Das, S. A. Shammi, and N. R. Chakraborty, "Monovab: An annotated corpus for bangla multi-label emotion detection," *arXiv preprint arXiv:2309.15670*, 2023.

[4] Z. Guo, A. Lai, J. H. Thygesen, J. Farrington, T. Keen, and K. Li, "Large language models for mental health applications: Systematic review (preprint)," Feb. 2024. DOI: 10.2196/preprints.57400 [Online]. Available: http://dx.doi.org/10.2196/preprints.57400

[5] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, pp. 1735–1780, Nov. 1997. DOI: 10.1162/neco.1997.9.8.1735

[6] E. J. Hu et al., "Lora: Low-rank adaptation of large language models.," *ICLR*, vol. 1, no. 2, p. 3, 2022.

[7] M. A. Iqbal, A. Das, O. Sharif, M. M. Hoque, and I. H. Sarker, "Bemoc: A corpus for identifying emotion in bengali texts," *SN Computer Science*, vol. 3, no. 2, p. 135, 2022.

[8] M. S. Islam, M. E. Rahman, M. S. Moonajilin, and J. van Os, "Prevalence of depression, anxiety and associated factors among school going adolescents in bangladesh: Findings from a cross-sectional study," *Plos one*, vol. 16, no. 4, e0247898, 2021.

[9] M. S. Islam, M. M. Islam, M. R. Islam, M. S. Hossain, and M. R. Islam, "Mental health status among university entrance test-taking students in bangladesh: A cross-sectional study," *Scientific Reports*, vol. 14, no. 1, pp. 1–10, 2024, Accessed:

2025-04-26. DOI: 10.1038/s41598-024-72235-z [Online]. Available: https://www.nature.com/articles/s41598-024-72235-z

[10] S. Islam et al., *A comprehensive survey on applications of transformers for deep learning tasks*, 2023. arXiv: 2306.07303 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2306.07303

[11] J.-W. Kim, J.-E. Han, J.-S. Koh, H.-T. Seo, and D.-S. Chang, *Enhancing psychotherapy counseling: A data augmentation pipeline leveraging large language models for counseling conversations*, 2024. arXiv: 2406.08718 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2406.08718

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

[13] W. H. Organization. "Minister of health releases first findings of national mental health survey." Accessed: 2025-04-26. [Online]. Available: https://www.who.int/bangladesh/news/detail/27-11-2019-minister-of-health-releases-first-findings-of-national-mental-health-survey

[14] D. Pant and S. K. Mukhiya, "Technical evaluation of gpt-4 and gpt-4-turbo reflective listening response generation with the counsel chat dataset," 2024.

[15] S. Patil and A. Rasave, "Artificial intelligence chat bot for counselling therapy," in *Proceedings of the 4th International Conference on Advances in Science Technology (ICAST2021)*, 2021.

[16] A. J. Trappey, A. P. Lin, K. Y. Hsu, C. V. Trappey, and K. L. Tu, "Development of an empathy-centric counseling chatbot system capable of sentimental dialogue analysis," *Processes*, vol. 10, no. 5, p. 930, 2022.

[17] A. Vaswani et al., *Attention is all you need*, 2023. arXiv: 1706.03762 [cs.CL]. [Online]. Available: https://arxiv.org/abs/1706.03762

[18] X. Xu et al., "Mental-llm: Leveraging large language models for mental health prediction via online text data," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 8, no. 1, pp. 1–32, Mar. 2024, ISSN: 2474-9567. DOI: 10.1145/3643540 [Online]. Available: http://dx.doi.org/10.1145/3643540

[19] X. Zhang and Z. Luo, *Advancing conversational psychotherapy: Integrating privacy, dual-memory, and domain expertise with large language models*, 2024. arXiv: 2412.02987 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2412.02987

Appendices

Appendix A

VRAM requirement for complete finetuning vs LoRA finetuning

Compared to full fine-tuning, which requires modifying all model parameters and hence demands significantly higher VRAM, LoRA (Low-Rank Adaptation) fine-tuning drastically reduces memory usage by introducing a small number of additional trainable parameters. As shown in Table A.1, full fine-tuning typically requires 12–24 GB or more of VRAM, making it resource-intensive and often impractical without high-end hardware. In contrast, LoRA fine-tuning can operate efficiently within 2–6 GB of VRAM, allowing model customization even on more modest computational setups. This substantial reduction in memory requirements makes LoRA a highly attractive alternative for efficient, scalable model adaptation without sacrificing much performance.

Table A.1: Comparison of VRAM and Resource Requirements: Full Fine-Tuning vs. LoRA Fine-Tuning

Aspect	**Full Fine-Tuning**	**LoRA Fine-Tuning**
Trainable Parameters	100% of parameters	∼0.1%–1% of parameters
VRAM Requirement	12–24 GB+	2–6 GB
Training Speed	Slower	Faster
Hardware Needed	High-end GPUs	Consumer-grade GPUs
Flexibility	Less flexible	Highly modular