Developing an Empathetic Conversational System using Fine-tuned Language Models and Psychometric Evaluation

by

Shuha Jahan
21301335
Rifa Tasnim
21301565
S.M Sajidur Rahman
21301130
Mashira Rofi
21301169

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science and Engineering

Department of Computer Science and Engineering
Brac University
December 2025

Declaration

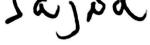
It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Shuha Jahan
21301335

Rifa Tasnim
21301565



S.M Sajidur Rahman
21301130

Mashira Rofi
21301169

Approval

The thesis/project titled "Developing an Empathetic Conversational System using Fine-tuned Language Models and Psychometric Evaluation" submitted by

1. Shuha Jahan (21301335)
2. Rifa Tasnim (21301565)
3. S.M Sajidur Rahman (21301130)
4. Mashira Rofi (21301169)

Of Fall, 2025 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science in 2025.

Examining Committee:

Supervisor:
(Member)

Md. Aquib Azmain

Md. Aquib Azmain

Lecturer
Department of Computer Science and Engineering
Brac University

Thesis Coordinator:
(Member)

Dr. Md. Golam Rabiul Alam

Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Dr. Sadia Hamid Kazi

Associate Professor & Chairperson
Department of Computer Science and Engineering
Brac University

Abstract

Developing an empathetic conversational system remains challenging, particularly when models must generate contextually relevant, accurate, and sensitive responses from limited data. Recent advances in large language models (LLMs) have demonstrated that architectural innovation strategies can enhance dialogue performance and adaptability. In this work, we experimented with multiple LLM architectures: Gemma 2 9B IT, Zephyr 7B Beta, Llama 2 7B and Phi-3 Mini 4K Instruct for dialogue generation and structured question–answer tasks, and selected the best-performing model based on response quality and accuracy.

The models were evaluated on a dataset of 839 samples, derived from validated psychometric scales including MSPSS (Multidimensional Scale of Perceived Social Support), PHQ-4 (Patient Health Questionnaire-4), and PSS-4 (Perceived Stress Scale-4), curated and reviewed by a psychiatrist to ensure clinical relevance. MSPSS measures perceived social support from family, friends, and significant others; PHQ-4 screens for anxiety and depression severity; and PSS-4 assesses perceived stress levels, with reverse scoring for certain items. Responses to these scales were converted into numeric scores and integrated as structured input, allowing the selected model to generate dialogue that adapts to each user's emotional state, stress level, and social support network. Parameter-efficient fine-tuning techniques such as LoRA and PEFT were employed to maximize learning from the small dataset.

Overall, this study demonstrates that instruction-tuning a carefully chosen LLM with psychometric-informed inputs enables empathetic, context-aware, and clinically informed dialogue, offering a promising approach for personalized mental health assessment and conversational support.

Keywords: MH, Mental Health, Support, Bot, Chatbot, Language Model, LLM, Large Language Model, Instruction-Tuned, Low-Rank Adaptation (LoRA), Psychometric Scales, MSPSS, PHQ-4, PSS-4, Gemma 2 9B IT, Zephyr 7B Beta, Llama 2 7B and Phi-3 Mini 4K Instruct

Acknowledgement

Firstly, all praise to the Great Allah for whom our thesis have been completed without any major interruption.

Secondly, to our respected advisor, Md. Aquib Azmain sir for his kind support and advice in our work. He helped us whenever we needed help.

We Are also greatful to Dr. Atiqul Haq Mazumder, MBBS, MD&PhD(Psychiatry) for his kind and constant support throughout his busy schedule.

And finally to our parents without their throughout sup-port it may not be possible. With their kind support and prayer we are now on the verge of our graduation.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Introduction

Mental health dysfunctions compose a significant public health threat worldwide, with principally harsh hypotheses for low and middle income countries (LMICs). The nation comes across an enormous accountability of sickness from both communicable and non-communicable diseases, which also include mental concerns, yet medical care for mental health stays substantially deficient because of finite public services, lack of certified professionals and social stigma. As stated by the National Mental Health Survey in 2019, about 19% of the population, mostly adults, in Bangladesh suffers from some kind of mental health issue[20], exhibiting a significant rise from the 16.1% prevalence studied in the survey of 2003-2005[12]. The pandemic due to corona virus has heightened this crisis even more, creating many mental health difficulties among all demographic groups[34].

Regardless of this increasing burden, it was reported by the 2018 to 2019 Bangladesh National Mental Health Survey that approximately 92.3% of the people of this nation with identifiable mental illness were not obtaining proper psychological aid[47]. This concerning gap in mental health treatment steps from multiple aspects: Bangladesh, with a population of 162 million, has only 565 and 260 psychologists and psychiatrists respectively to serve the entire nation, with maximum of these professionals in urban areas[52]. Furthermore, only 0.44% of the country's entire mental health budget is being spent, causing major issues in delivering proper services for those with mental health conditions. Factors like societal diffident, limited awareness and disclosure of personal information add to these challenges, stopping people from asking for help even when facilities are available[53]

Large Language Models (LLMs) and Artificial Intelligence (AI) have surfaced as effective solutions to tackle the nationwide mental health service gap. Chatbot-based systems have exhibited potential to deliver support for people with psychological and social disabilities, when direct human engagement is not accessible or wanted. LLM-based support systems can implement advanced techniques like instruction tuning, fine-tuning and much more, allowing them to offer culturally relevant assistance in mental health [51]. Specifically, for the students of Bangladesh, chat-based bots facilitate a route to express personal psychological matters without being conscious about criticism or societal judgement which usually act as an obstacle from

looking for traditional care[42].

This study establishes a chatbot for psychosocial support particularly constructed for the English-speaking people of Bangladesh, employing instruction-tuning on models with culturally relevant dataset. The framework combines clinically validated assessment tools— Patient Health Questionnaire (phq4), Multidimensional Perceived Social Support Scale (MSPSS), and Perceived Stress Scale (PSS4) to make diagnosis and screening with proper evidence more available. With the integration of psychometric scales in conversational models, this tool intends to build something which is obtainable, easy-to-use and aligned with cultural context , and that can also aid to minimize the mental health treatment gap in the country.

1.2 Problem Statement

Bangladesh comes face-to-face with a wide range of serious mental health challenges, with high prevalence, unmet gaps in treatment, and significant barriers in levels of the system in terms of care. The severity of this crisis is reflected through multiple indicators:

Critical Condition and Low Resources: Almost 1/5th of the entire population of Bangladesh is affected by mental health conditions, 6.7% with depression, 4.7% with anxiety and somatic symptoms in 2.3% being very widespread. With the population aged between 7 to 17 years, mental illness is experienced by nearly 13.6%. Nonetheless, support for mental issues is still lacking in the healthcare infrastructure[52].

Huge Gap in Treatment: Bangladesh's mental health situation is marked by a huge gap between people who need care and the people who actually get that care. Research highlights that treatment-seeking among people is very low. However, this gap remains even with so many disorders, suggesting that the availability of resources alone is not enough deeper barriers like structural, social and financial aspects should also be looked into[43].

Societal and Cultural Obstacles: Conditions and disorders related to mental health are way too overlooked and most of the time misunderstood as supernatural cases. This consciousness of being judged by others, being bullied or discriminated, made fun of, unemployment and even loss in marriage stop people from seeking treatment. These challenges are greater for women and younger people, because of having low decision making power in terms of healthcare choices[41].

Local and Linguistic Relevance: Existing mental health conversational bots are often for the English-speaking western population. While the use of English is increasing day-by-day in Bangladesh. Especially within the educated urban community, there is a scarcity of AI bots with the proper cultural context of Bangladesh.

Psychometric Scales Integration: There are lots of AI chatbots that focus only on conversational support but do not combine validated assessment tools to prop-

erly understand the level of support an individual needs. Even though some bots do integrate these tools, still it is limited, hence there is a high need for such systems to integrate MH analysis techniques while providing emotional support and validating the individual's feelings.

Challenges of Open-Ended Response: There are options in the form of Likeart Scales when using traditional psychometric scales making it seem more like an exam rather than sharing feelings and struggles. But, mapping an individual' s open-ended responses to the options of the psychometric assessment tools' options become a grave challenge. Hence, NLP techniques are required to build a stable bridge between these, and allow a natural flow of communication while maintaining psychometric validation[44][50].

To tackle these challenges, this research aims to develop a psychosocial support conversational bot, fundamentally tailored for the cultural context of the people of Bangladesh, instruction tuned on proper prompt and locally relevant dataset collected via interviews with the nation's people, and merging psychometric scales with the help of NLP for its understanding capabilities.

1.3 Research Objectives

The initial goal of this thesis is to construct and analyze an AI-based conversational bot for mental health support, modified for the English-speaking population of Bangladesh, which is competent to give emotional replies, simultaneously taking assessments through validated questionnaires. The primary objectives of this thesis are:

Objective 1: Constructing a Culturally Appropriate Dataset

- Collect conversational English interviews from the people of Bangladesh capturing different communication styles, mental health issues and context. If interviews are in Bangla, translate to english.
- Make sure the dataset has diversity in emotions and help-seeking behaviours
- Process the dataset for instruction-tuning of LLMs.

Objective 2: LLMs for Instruction Tuning

- Use pre-trained models and implement instruction-tuning with the usage of the collected bangladeshi dataset.
- Test different models to find the most optimal hyperparameters and architectures.
- Evaluate the performance of the models on the quality of conversation and proper. cultural nuances and relevant replies to users.

Objective 3: Integration of Psychometric Tools

- Combine 3 clinically validated MH analysis exams into the conversatiuonak system:
  - PSS-4
  - MSPSS
  - PHQ-4

Objective 4: Map Open-Ended Input Using NLP

- Use NLP similarity model to map the free-text reponses of users to the correct options of the psychometric scales.
- Keep the natural conversation flowing while assessing the user with the aid of tools
- Test how reliable the mapping and accuracies are

Objective 5: Merge the Instruction-Tuned Model and Assessment Tools

- Create a workflow where the warmup session is of 3-4 turns and then psychometric analysis starts.
- Ask follow up questions while using psychometric scales

Objective 6: Measure Performance and Clinical Impact

- Evaluate the performance of chatbot on:
  - Quality of conversation and its engagement with users
  - NLP mapping accuracy
  - Cultural relevance and user completion and satisfaction rate

By attaining all these objectives, the study is eager to contribute to the mental health infrastructure of Bangladesh and diminish those treatment gaps. The final system shows a novel integration of clinically verified MH instruments with instruction-tuned models, and mapping techniques with NLP especially made for the Bangladeshi context.

Chapter 2

Literature Review

2.1 Mental Health Landscape in Bangladesh

Bangladesh is experiencing a growing mental health crisis specifically due to the limited resources stigma spread widely across the country. [12] reported that between 6.5% and 31% of the adults suffer from mental disorders, that includes depression and anxiety. The rates are even higher among students enrolled in universities, and over half of them were reported having symptoms of depression and anxiety[28].

The World Health Organization's Mental Health Atlas[35] showed that Bangladesh has only 0.49 psychiatrists for every 100,000 people, which is below the global average. Mental health services are found mostly in urban areas, leaving a majority of the rural communities with little or no access to professional care[28].

Furthermore, cultural stigma makes it even more difficult for people to seek help. Mental illness is often perceived as a sign of weakness, family shame, or at times a supernatural punishment[12]. Consequently, many individuals turn to traditional healers even before they consult medical professionals to help them [12].

2.2 Conversational AI in Mental Healthcare

AI chatbots are becoming an interesting option for mental health support, especially in places with few therapists. Early systems like ELIZA could mimic conversation but didn't really "understand" what users were feeling [2]. Modern tools, such as Woebot and Wysa, are much smarter. They combine therapy techniques with AI to help people manage anxiety or depression. Interestingly, studies show they can improve mood and keep users engaged, particularly young adults [15][17].

That said, AI isn't a complete replacement for humans. Privacy, ethical concerns, and handling crisis situations remain tricky. Experts suggest using AI alongside human supervision for safety [26].

2.3 Large Language Models and Mental Health Applications

Have you ever spoken with a chatbot that sounded like it knew how you feel? Large language models (LLMs) like BERT and GPT-3 have been trained to do the same. They use transformer architectures, so they can read the whole conversation at once rather than sentence by sentence [16][19]. Some are fine-tuned specifically for mental health. For example, DialoGPT was trained on therapy sessions, and the researchers discovered that nearly half of its responses were perceived as truly empathetic [27][32]. Just type out your concerns and receive a reply that seems as though someone who truly gets it is speaking; it won't be perfect, perhaps a little more so.

Instruction-tuned models like InstructGPT and Gemma 2 can better follow instructions and even consider cultural differences. However, they can commit mistakes or ignore crucial warning signs, so it is essential to have a human oversee them [37].

2.4 Parameter-Efficient Fine-Tuning (PEFT) and LoRA

Retraining from scratch all large language models is extremely slow and expensive. Luckily, there is a more efficient method called Parameter-Efficient Fine-Tuning, or PEFT. Scientists don't retrain the whole model but fine-tune only a tiny fraction of it. It is like adding some new tools to a well-stocked toolbox without replacing them[36].

One of the widely used PEFT methods is LoRA, where small learnable matrices are added to the model while the original weights are frozen. This enables efficient learning of new tasks without forgetting what has already been learned. QLoRA takes this a step further by quantizing the model to fewer bits and hence even enormous models can be fine-tuned on ordinary computers [40].

This approach is especially useful for mental health chatbots. For example, a model may be fine-tuned to acquire local terms for language or culture using just a very small dataset. That makes it most applicable to countries such as Bangladesh where there is little computing power.

2.5 Psychometric Assessment in Mental Health

Systematic evaluation of mental health symptoms is made possible by standardizing psychometric tools, which offers trustworthy and validated assessments of psychological notions. In large-scale population studies and situations with limited resources, where brief screening tools are very helpful but thorough clinical interviews are not feasible.

Patient Health Questionnaire-4 (PHQ-4): PHQ-4, or the Patient Health Questionnaire,[10] created the ultra-brief PHQ-4, a screening tool that combines two items from the GAD-2 anxiety screener and two things from the PHQ-2 depression screener. In just two weeks, this 4-item assessment effectively examines the fundamental symptoms of anxiety and depression. A two-factor structure (anxiety and depression) accounts for 84% of the variance, according to factor analysis, and there are strong correlations between rising PHQ-4 scores and functional impairment, disability days, and healthcare use.

Perceived Stress Scale-4 (PSS-4): Based on the transactional distress model,[4] Cohen and associates created the first Perceived Stress Scale to measure subjective evaluations of life stressors. Though there are 10-item (PSS-10) and 14-item (PSS-14) variations, the 4-item simplified version has the least amount of responder burden. However, studies show that PSS-4 has irregular internal consistency across populations and worse psychometric qualities than the extended variants.However, the PSS-4 is still utilized when any extreme clarity is required , and this is adapted for many cultural concepts.

Multidimensional Scale of Perceived Social Support (MSPSS): [5] From three different sources a 12-item self-report test that measures the perceived sufficiency of social support: friends (4 items), family (4 items), and significant others (4 items). With Cronbach's alpha generally falling between 0.95 and 0.85 across a range of populations and cultural scenarios, the scale displays outstanding internal stability. Similar robustness is seen in test-retest reliability, with intraclass correlation values typically above 0.84. Confirmatory factor analysis has been applied in numerous investigations to successfully validate the three-layer structure.in numerous investigations to successfully validate the three-layer structure.

2.6 NLP Model for Psychometric Response Assessment

The all-MiniLM-L6-v2 model that has been used from the Sentence-Transformers library is utilized by the psychometric assessment system to map natural language answers and standardize psychometric scale possibilities,[22] .It's engaged with six transformer layers, 384-dimensional embeddings, and roughly 22.7 million parameters; this model is a condensed form of BERT that uses deep attention distillation.

The approach uses mean pooling, tokenization and multi-layer transformer encoding to convert user text into viscous semantic vectors. The three steps of response are mentioned here where the first one is - the users natural language response is encoded into a 384-dimensional vector; second, several linguistic variations of each scale option are encoded (for example, "nearly every day" includes variations like "always," "constantly," and "daily"); and the best is third one, which is semantic match determined by computing the cosine similarity between the user embedding and all option embeddings.[30].

Cosine similarity looks at the angle between vectors in high-dimensional space and gives scores from 0 to 1,[25], with larger scores meaning that the vectors are more semantically oriented. And for response classification, a threshold of 0.55 guarantees moderate-to-high confidence. This method works especially well in case of assessments like naturalistic psychometric scenarios because it allows the system to accurately achieve paraphrased and causal comments.

2.7 Instruction-Tuning and Fine-Tuning

Fine-tuning modifies previously learned model parameters; instruction tuning is a particular variation that improves models' comprehension of natural language commands. Targeted training approaches are necessary to adapt pre-trained language models to specialized domains, not withstanding their excellent capabilities. In order to modify model parameters for better performance on specific applications, fine-tune them and carry on with the training process using task-specific data [19].

A specialized type of fine-tuning called instruction tuning improves models' comprehension, their ability to interpret and performance of natural language commands by training them on a variety of instruction following tasks[39]. Instruction tuning in essence, is fine-tuning with an emphasis on enhancing task knowledge, thoughtfulness, empathy, and clarity using natural language instructions. These models are further personalized to specialized situations, such as mental health counseling, through domain-specific fine-tuning, and this happens while it's built upon an instruction-tuning basis.

There are two main strategies: efficient parameter techniques like [36] Low-Rank Adaptation (LoRA), which trains a limited selection of adapter weights while freezing the base model, and full fine-tuning, which updates all the model parameters. And both these strategies are still underutilized for counseling implementations, especially when it comes to comparative research that looks at how well they produce compassionate, sympathetic, contextually relevant reactions in settings with limited resources.

2.8 PostHoc and CoT

Post-hoc is a statistical analysis which is a highly useful and flexible technique to determine what LLM models are actually implementing. This test can easily be applied to any pre-trained model to see what they are doing without modifying its architecture. Feature attribution, attention visualization and saliency maps are widely used by researchers to rapidly detect which part of the input is prioritized by the model's output. For cases like biases, debugging and explaining decisions (i.e. finance or healthcare), post-hoc is very valuable as it can build a sense of trust after providing proper explanations after facts.[23][29]

Another simple but very high-powered method that enhances the reasoning abilities of LLM models is CoT, Chain-of-Thought. On challenging tasks, CoT can be implemented to check how the model explains what it did. This makes LLM much better at complex problems like symbolic manipulation, iterative arithmetic and logical thinking. CoT makes sure to naturally output reasoning traces, by giving a step-by-step guide, which shows how the model came to that conclusion making it simpler for developers to verify. This mixture of built-in understandability and notable increase in performance has made CoT one of the most fundamental and efficient promoting strategies in LLM.[39][38]

Chapter 3

Methadology

3.1 System Workflow

This chapter describes an extensive overview of the workflow designed for the implementation of the proposed mental health system. The workflow consists of dataset collection and curation, validation of experts, dataset preprocessing, model training, integration of clinical tools, and mapping of free-texts using natural language processing (NLP) methods.

This research started with the construction of a professional verified dataset encompassing interviews, natural conversations, audio recordings and surveys, These data sources guaranteed a variety of linguistic representations, ecological validity and emotional depth. The dataset was then used to instruction-tune multiple preprained large language models (LLMs), Llama 2 7B Chat, Zephyr 7B Beta, Phi-3-Mini 4K Instruct and Gemma 2 8B IT. One single standard instruction was used, during tuning, which was reviewed by a certified professional. This prompt was applied consistently to all the models so the training remains uniform and decreases instruction-based imbalance in the behaviours of the models.

After the completion of model instruction-tuning, the performance of each model was evaluated using multiple different evaluation metrics – Semantic Similarity, BERTScore F1, ROUGE. The best performing model was lastly used for the system.

For clinical validation, the model was advanced by merging three globally validated psychometric scales: the Perceived Stress Scale-4 (PSS-4), the Multidimensional Scale of Perceived Social Support (MSPSS), and the Patient Health Questionnaire-4 (PHQ-4). These scales are known to be reliable and accurate for diagnostic validity across different populations. To allow open-ended replies to scales options, instead of users having to choose predefined multiple choice options, the thesis used a NLP-based model for mapping, This mapping layer uses a transformer based similarity model and a database of synonyms, which include 30-40 variations for each option of the scales' items. This technique enabled the system to correctly understand the user input and map it to the accurate option of the scale.

The final system thus combines:

1. An instruction-tuned LLM using a validated dataset.

2. Psychometric assessment tools for clinical validation

3. NLP semantic similarity model for free text responses to scales options

Overall, the combination gives empathetic replies, builds trust and validate individuals' feelings and help them to share their struggles.

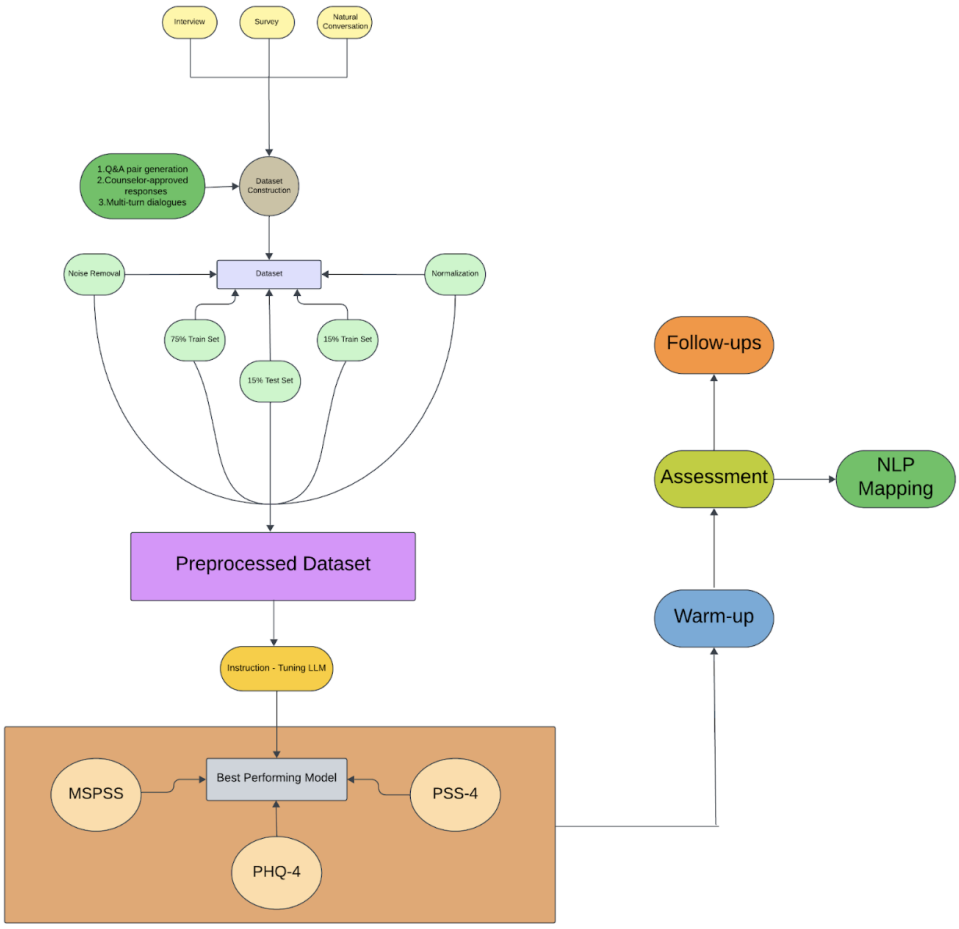


Figure 3.1: Study Workflow

3.2 Data Collection and Curation

3.2.1 Informed Consent

All individuals taking part in the dataset collection process were informed about the reason, process, and consequences of the study before their involvement. The interviewees were given a briefing on the nature of the interviews, and how their responses would be used for academic research. They were also informed regarding the privacy concerns and their right to end the interview at any time. Based on the preference of the participants, a verbal/written consent was taken. This method of dataset collection upheld both voluntary participation and any other ethical concerns.

3.2.2 Participant Recruitment

The interviews conducted for data collection have a variety of participants across Bangladesh. It included people from urban and rural parts of the nation, and a small group of Bangladeshis living out of the country. For a diverse presentation, participants of different gender, age and status were recruited for an interview. Inclusion criteria were:

- Citizens of Bangladesh
- Willing to take part in this research to talk about their mental health.
- Able to talk in either English, or Bengali.

Demographic Variables	Percentage(%)
Gender: Male	~ 46
Gender: Female	~ 54
Age: 0-14 years	~ 22
Age: 15-64 years	~ 22
Age: 65+ years	~ 7
Occupation: Student/Employed	~ 59
Occupation: Unemployed	~ 13
Occupation: Other	~ 28
Urban Population	~ 68
Rural Population	~ 32

Table 3.1: Participant Demographics

3.2.3 Data Collection Procedure

The primary way of collecting data was through semi-structured interviews. The protocols of interview were collectively constructed and judged by a professional to guarantee clinical relevance. The interviews were carefully designed but flexible too, allowing the interviewees to share their feelings naturally. The semi-structured interviews were taken in the following ways:

- Verbally consented audio recordings
- Manually written responses
- Survey based inputs

Some interviewees were comfortable using the Bengali language and hence their responses were translated to English while making sure to preserve the meaning and emotional tone for proper study.

Interview flow:

1. **Introduction and Rapport (1-2 minutes)**
   The researchers started the conversations/ interviews with a small introduction to build trust and comfort.

2. **Free-Response Questions (2-5 minutes)**
   The questions asked to the participants were as follows:

   - "How have you been feeling these days?"
   - "What makes you feel down, stressed or anxious?"
   - "Who or what makes you feel calm and better during troubled times?"

   These questions were not only discussed by the researchers but they were reviewed by a psychiatrist as well, who stated that the prompts enable users to share their thoughts and experiences in their own ways. In this stage, stressors and triggers like academic stress, work life concerns, family conflicts, and health issues were also discussed.

3. **Professional Verification** The labels given to each participant after their interview was collectively checked by the psychiatrist to ensure the validity, and proper and correct categorization.

3.2.4 Dataset Composition

The final dataset included a total of 839 conversational pairs. Each entry consisted of a message from the user articulating and emotional or mental issue and an emotionally validating system reply. Examples of inputs of the users ranged from basic distress ("Tensed these days") to specific emotional turmoils ("Felling stressed about tomorrow's office meeting").

Each entry included:

- **User Input:** An individual's expression of mental health concern, ranging from general distress ("Been feeling too overwhelmed lately") to particular incidents ("My father is very sick and we are also not very financially stable at this moment")
- **Bot Reply:** Empathetic, supportive and clinically validated response


```
{
  "id": 10,
  "user_input1": "Feeling heavy...",
  "system_response1": "I hear how...",
  "user_input2": "My Aunt's...."
  "system_response2": "Losing you...",
  "user_input3": "Praying with...",
  "system_response3" : "Praying with..."
  "category": "Life-event stress"
},
{
  "id": 11,
  "user_input1": "I am very...",
  "system_response1": "I'm so sorry...",
  "user_input2": "Commuting...",
  "system_response2": "Daily commute...",
```

```
  "user_input3": "Spending...",
  "system_response3": "It's wonderful...."
  "category": "Social stress"
},
{
  "id": 12,
  "user_input1": "Drained from...",
  "system_response1": "It sounds...",
  "user_input2": "The Fieldwork...",
  "system_response2": "Fieldwork...",
  "user_input3" : "Sharing story...",
  "system_response3" : "Sharing stories..."
  "category": "Work Pressure"
}
```


Dataset Statistics:

- Total interactions: 839
- Average length of user input:29.4008 tokens ($\sigma = 15.0267$)
- Average length of bot reply: 101.7675 tokens ($\sigma = 16.0230$)
- Vocabulary size: 6,720 unique tokens
- Topics covered: Personal Stress (12%), Academic Stress (11%), No Stress/Positive State (11%), Family Issues (10%), Work Pressure (10%), Health Concerns (9%), Career Stress (9%),etc.

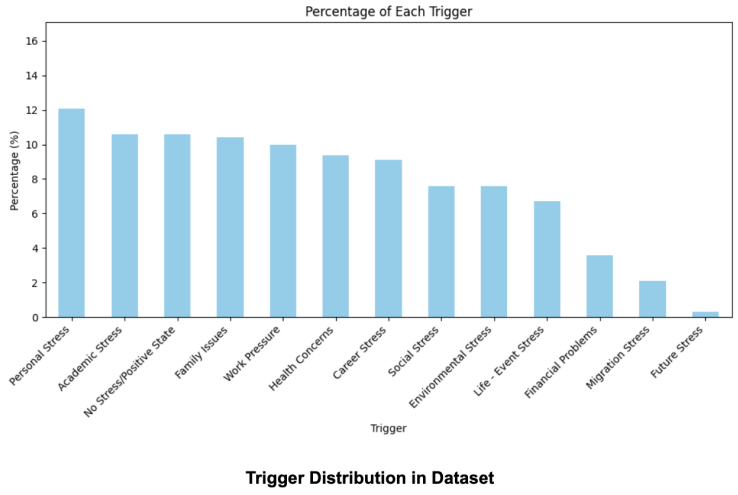


Figure 3.2: Distribution in Dataset

3.2.5 Clinical Validation

The entire bot responses from all the 3 rows (R1,R2,R3) were formulated adhering evidence-based standards from:

- Person-Centered Therapy[1]: Respectful acceptance, compassionate understanding and authenticity
- Motivational Interviewing [11]: Free-form question, validating comments and summarizing key points

Dr. Atiqul Haq Mazumder (MBBS, MD&PhD(Psychiatry)), a licensed psychiatrist, with over 20 years of experience in mental healthcare, carefully judged the entirety of the dataset to certify the alignment with mental health standards, non-directive advice and vigilant language use.

3.3 Dataset Preprocessing

3.3.1 Text Cleaning

Before the model tuning, the raw dataset was passed through a multi-step normalization and cleaning process:

1. Noise Removal:
   (a) Deletion of filler words ("actually", "hmm", "ok", "ah")
   (b) Elimination of same words
   (c) Fixing incomplete words/phrases
2. Normalization:
   (a) Converting short forms to full forms ("btw" → "by the way")
   (b) Correcting misspelled words/phrases
   (c) Fixing punctuation errors and white spaces.

These preprocessing steps helped to not negatively affect the tuning of the model and that the meaning of the dataset remained uniform across multiple sources of data.

3.3.2 Train-Validation-Test Split

A stratified split was implemented to maintain the balance of the theme during model tuning. This split was performed based on the verified "Trigger" column of our dataset. This technique helped to maintain topic distribution:

1. Training Set: 587 QnA pairs (70%)
2. Test Set: 126 QnA pairs (15%)
3. Validation Set: 126 QnA pairs (15%)

Sampling bias was reduced through this technique and it supported the models for the generalization of the diverse MH storylines.

3.4 Model Selection Criteria

The specifics of the experiment in the paper required the careful choice of big language models that can provide psychosocial support as empathy and situational appropriateness. The selection of models was done with regard to a number of important factors. To begin with, every candidate had to use the parameter-efficient fine-tuning techniques because resource limitations are a crucial factor in creating available mental-health technologies that can be used in Bangladesh. Secondly, the models were to exhibit high instruction-following skills, which are important in case of responding to sensitive user queries in a counseling setup. Third, the strong focus was made on the efficiency of the computations to be performed to make sure that the end system would be able to run on relatively easy to obtain hardware instead of having to install costly infrastructure. Finally, all the chosen models are open-source and, thus, ensure transparency and reproducibility of our research both in addition to the ethical considerations of the open healthcare technology[36].

The list contains four model families, each one of which is a different way of modeling language. The conversational baseline was chosen as Zephyr 7B Beta which was specifically trained in dialogue using Direct Preference Optimization[45]. Gemma 2 9B Instruction-Tuned represents a current interest in Google in the safety-consistent models with superior instruction-following[49]. The flagship product of meta, 7B Instruct Llama 2 7B was integrated as a high-performance reference due to its high overall general performance[48]. Last but not least, Phi-3 Mini 4k Instruct was tested to show whether small architectures can be used to obtain a decent performance even with their small size[46]. The wide range of possibilities allows a strict comparison of the model scales and architectural philosophies, but does not exceed the realistic computational possibilities.

3.5 Evaluation Metrics and Target Thresholds

Psychosocial support chatbots need to be evaluated using measures that transcend the traditional language modeling metrics. Mental health apps require tight semantic correspondence to best practices in therapy and empathetic intonation, unlike the general conversational AI. Similar to these needs we put up a three-layer assessment model. The main measure is Semantic Similarity which measures the resemblance between responses generated by the model and the gold-standard counseling responses in semantic space. This measure is based on sentence-transformer embeddings since it captures deep meaning of words more than overlap on the surface. To determine that a model is fit to use in counseling practice, the semantic similarity should be greater than 0.60 because anything below this will be taken to mean that the responses are not therapeutically intentional or they lack proper guidance. BERTScore F1 is used to derive contextual quality by assessing semantic similarity at token-levels by using contextualized embeddings in pretrained BERT models. This measure will prevent responses that are not coherent and of the right context all the way through with a desired value of 0.80 denoting a good quality of contextual match. Perplexity is used to measure the confidence of the model to produce responses and low values show more natural and fluent language generation. In counseling applications, where the trust of the user is the most important, per-

plexity should not exceed 5.0 to guarantee answers are not artificial and doubtful. ROUGE-L uses structural alignment, based on the longest common subsequence between generated and reference responses, to model the information flow and the coherence of an organization. The target of 0.30 makes the responses adhere to rational therapeutic patterns. Response Length Consistency monitors average word count and standard deviation, where ideal is 20 to 50 words denoting brief and full therapeutic answers.

ROUGE-1, ROUGE-2, BERTScore Precision, and BERTScore Recall were also calculated to have complete documentation since these additional measurements would provide information about a certain feature of the quality of responses. These metrics have been calculated to the established implementations to guarantee reliability and comparability. To compute semantic similarity, the sentence-transformers library was used on all-MiniLM-L6-v2 models[22]. The calculation of BERTScore was calculated with the bert-score package with the reference model of bert-base-uncased[24]. The standard rouge-score implementation was used to calculate ROUGE scores[8]. The loss of cross-entropy was computed using the held-out test data to come up with perplexity[13]. All measures are presented in mean values including standard deviation calculated in the entire test set. In order to make results reproducible, the random seed was fixed at 42 in all experiments, making stochastic variation in the initial weights and batch sampling deterministic.

3.6 Experimental Protocol

The experiments were done on a culturally adapted psychosocial support data that was developed specifically on the Bangladesh setting. The data covers mental health-related issues that are common among the local people and uses language and treatment methods that are culturally relevant to counseling.

3.6.1 Dataset and Training Configuration

The experiments were done on a culturally adapted psychosocial support data that was developed specifically on the Bangladesh setting. The data covers mental health-related issues that are common among the local people and uses language and treatment methods that are culturally relevant to counseling.

The full dataset consists of 839 samples which were divided in a manner that would be able to give strong model assessment and avoid overfitting. Data was divided in 70-15-15 (587 samples, 126 respectively, and 126 samples served as a held-out test set). Hyperparameter tuning and early stopping decisions were made using the validation set, but the test set was not correctly evaluated until the end of training to give an objective evaluation of model performance.

It is necessary to mention that the test sizes used in models differ slightly because of logistics in experiments. Zephyr, Gemma, and Llama models were tested on 126 test samples with Phi-3 Mini on another partition that had 84 samples. This difference does not affect the validity of within-model performance measures but must be taken into account during the direct cross-model comparisons. No scientific methods were

used to enrich the data of either the training or the test sets because the assessment had to be based on actual counseling situations and not the artificially improved distribution.

3.6.2 Evaluation Methodology

Since the computational demands of implementing accessible technologies in resource-limited situations, the fine-tuning of all models was done by Low-Rank Adaptation with 4-bit quantization. The parameter efficient method, designated as QLoRA, minimises the memory utilisation significantly and preserves model expressiveness by inserting little trainable adapter layers into the frozen base model weights[40].

LoRA configuration took rank of 16 and scaled alpha of 32, and focused on query and value projection layers across the transformer. The dropout rate used was 0.05 to avoid overfitting on the small training data. To quantize, 4-bit NormalFloat precision was used with a bfloat16 compute datatype and with double quantization on in order to balance the numerical stability with the memory efficiency. All fine-tuning experiments were trained on the 700-sample training split and hyperparameter tuned on the 150-sample validation split. A patience of three epochs was used on early stopping and validation loss was checked to prevent overfitting. The training hyperparameters were made consistent across the models to remove architectural variations, but not training artifacts.

Each model had a batch size of 4 with a 4-step gradient accumulation effectively simulating a batch size of 16 and keeping within memory limits. The learning rate was also 2x 10-4 and 100 warm up steps and linear decay. To stabilize training, the AdamW optimizer was set to use default beta values of 0.9 and 0.999, weight decay of 0.01 and gradient clipping with the norm of 1.0 to stabilize training. Bfloat16 mixed precision training was also made available when hardware supported it.

Model	Type	Training Domain	Role in Study	Epochs	Batch Size	Learning Rate
Llama 2 7B	Conversational	Multi-domain dialogue	High-performance reference	3	4	2×10^{-4}
Gemma 2 9B IT	Instruction-tuned	Safety-focused instruction	Minimum training ablation	2	4	2×10^{-4}
Zephyr 7B Beta	Conversational	Multi-domain dialogue	Dialogue-optimized	2	4	2×10^{-4}
Phi-3 Mini 4K Instruct	Instruction-tuned	Safety-focused instruction	Efficiency oriented evaluation	2	4	2×10^{-4}

Table 3.2: Fine-Tuning Hyperparameters Across All Models

Time spent in training was the major difference among experiments. Zephyr 7B Beta was trained using 2 epochs, which is a conservative training regime. Both Llama 2 7B Instruct and Gemma 2 9B IT were trained on 3-epoch training proceeds, and Phi-3 Mini 4k was trained on 2-epoch training proceeds. In order to examine how training time influences quality of counseling, Gemma 2 9B IT was tested in two conditions; a minimal 2-epoch fine-tuning regime and a lengthy 3-epoch regime, which gives a controlled ablation study of convergence behavior.

3.6.3 Model Integration and Inference Protocol

The evaluation pipeline was created to test the performance of the models in the conditions of real-life counseling. In each sample of the test, a user query in the gold-standard dataset was given to the models, and responses were produced by nucleus sampling with a top-p value of 0.9 and temperature of 0.7. This arrangement moderates inventive variation, as well as consistent yield, preventing excessive repetition of responses as well as genocentric production. Each response was limited to a maximum of 150 new tokens with a penalty of 1.1 repeated tokens to prevent redundant phrasing. These parameters of generation were kept the same in all the models so that variations in evaluation measures would indicate real capabilities of the models and not the decoding strategy artifacts.

Simulation of resource-restricted healthcare environments All experiments were run on Tesla P100 and Tesla T4 GPUs with 4-bit quantization to simulate deployment conditions characteristic of resource-restricted healthcare environments. Peft and bitsandbytes packages were applied together with the Hugging Face transformers library to allow efficient training[31]. Such an experimental design forms a strict basis of comparing various architectures of language models on the domain specific task of psychosocial support, so that performance differences found would be due to real architectural properties and not due to confounding influences of the experiment.

Chapter 4

Results and Analysis

In this chapter, a deep analysis of five large language model configurations that were fine-tuned on psychosocial support under the Bangladesh context is discussed. All models were compared to the three tier model framework set in Chapter 6, with specific analysis of the key metrics that decide the readiness of the deployment BERTScore F1, Semantic Similarity, and Perplexity. Findings will be grouped in terms of individual model performance and then they will be compared and recommendations on the deployment will be made.

4.1 Individual Model Performance

4.1.1 Zephyr 7B Beta (LoRa + 4bit Quantization+ 2 epochs)

Using LoRA fine-tuning with two epochs and 4-bit quantization for computational performance, the Zephyr 7B Beta model was assessed as a baseline for psychosocial support chatbot applications.

Significant shortcomings were found in a number of important performance indicators. The model's semantic similarity score of 0.2412 ($\pm$0.1364) was considerably below the required threshold of 0.60 for effective counseling applications. The responses' lack of considerable semantic congruence with expected therapeutic content indicates that the model struggles to capture the complex linguistic patterns necessary for psychosocial assistance. Although the model demonstrated excellent contextual comprehension with a BERTScore F1 of 0.8471, meeting the minimum criteria of 0.80, its perplexity score of 12.0336 suggested concerning issues with response naturalness and confidence. This heightened perplexity suggests that the model produces unclear or awkward language that may not instill confidence in customers seeking emotional support.With ROUGE-L at 0.0936 and ROUGE-1 at 0.1250, both substantially below acceptable norms, the structural similarity measures showed more flaws.

The responses were inconsistently deep and may have been too short for meaningful conversations, with an average response length of 20.1 words and a high variance ($\pm$14.3). The Zephyr 7B Beta model received a low performance grade and needs to be greatly improved before being used in psychosocial support contexts because it only passed one of the three critical measures (BERTScore F1).

Metric	Value
BERTScore F1	0.8471
Perplexity	12.0336
Semantic Similarity	0.2412
ROUGE-L	0.0936
ROUGE-1	0.1250
ROUGE-2	0.0187
BERTScore Precision	0.8497
BERTScore Recall	0.8448
AvgResponse Length	20.1

Table 4.1: Complete Evaluation Report for Optimized Zephyr 7B Beta

4.1.2 Gemma 2 9B IT (LoRa + 1 epoch + 4bit Quantization)

Among the models studied, the Gemma 2 9B instruction-tuned model with 4-bit quantization and LoRA fine-tuning for a single epoch showed the best overall performance for psychosocial support applications.

Notably, this configuration met the essential threshold of 0.60 with a semantic similarity score of 0.6218 (±0.1376), making it the sole model to pass this crucial parameter for counseling chatbots. This accomplishment shows that the model can produce counseling information that is appropriate for the setting and successfully represents the semantic substance of therapeutic reactions.The model showed remarkable understanding of conversational nuances and emotional context, which are critical for support interactions, with the highest BERTScore F1 of 0.8738 (±0.0203). Additionally, the model generated considerably more comprehensive responses with an average length of 46.4 words and reasonable consistency (±17.7 words), which is more appropriate for psychological assistance, in contrast to the shorter outputs of other models.

However, a significant defect was revealed by the model's perplexity score of 28.142, which is far greater than the target of 5.0 for natural answers. This increased uncertainty implies that although the model generates information that is semantically accurate, the language may sound somewhat robotic or too formal, possibly lacking the conversational flow required in therapeutic circumstances. The structural alignment measures showed improvement over other models, with ROUGE-L at 0.1885 and ROUGE-1 at 0.3011, even if they were still below ideal standards. This model received a "GOOD" overall evaluation despite the perplexity problem, passing two of the three important metrics.

4.1.3 Analysis of Semantic Similarity

Zephyr 7B Beta model significantly underperformed With a score of 0.24 (±0.14), showing replies that lack meaningful semantic alignment with appropriate counseling material. Although it improved to 0.41 (±0.14) after three epochs of training, the Gemma 2 9B IT model was still below the target threshold, indicating semanti-

cally inconsistent replies. Most notably, out of all the models examined, the Gemma 2 9B IT model trained with one epoch demonstrated the best semantic alignment with expected counseling replies, meeting the aim with a score of 0.62 (±0.14). This finding implies that while prolonged training may paradoxically decrease semantic generalization, a single training phase is adequate to capture therapeutic language patterns without overfitting. Only the 1-epoch Gemma 2 configuration satisfies the minimal deployment requirements since semantic similarity is the most crucial parameter for guaranteeing therapeutically suitable responses in psychosocial support applications.

Higher scores indicate better grasp of conversational context. zBERTScore F1 uses BERT embeddings to test contextual understanding and semantic coherence.With a BERTScore F1 of 0.85 (±0.02), the Zephyr 7B Beta model demonstrated strong contextual comprehension. At 0.86 (±0.02), the Gemma 2 9B IT model trained over three epochs had marginally increased performance, indicating improved context management capabilities. Most remarkably, of all the models assessed, the Gemma 2 9B IT model trained with 1 epoch received the best score of 0.87 (±0.02), indicating the most advanced contextual comprehension. All three models consistently performed above the 0.80 standard, indicating that the fine-tuning technique successfully maintained and improved the models' comprehension of the intricate emotional and situational context.

The model's confidence and naturalness in language generation are measured by perplexity; lower values indicate more assured and natural responses. Because hesitant or robotic language can erode user trust and the therapeutic efficacy of interactions, this parameter is crucial for psychological support applications.The performance of the three assessed models differed considerably. With a perplexity score of 12.03, which was far higher than the goal threshold of 5.0, the Zephyr 7B Beta model had strong uncertainty, which could show up as hesitant or poor language in its responses. On the other hand, the Gemma 2 9B IT model, which was trained over three epochs, produced a remarkable perplexity score of 3.64, which was much below the target threshold and demonstrated the most confident and natural language creation of all the models. Despite the model's superior semantic similarity performance, the Gemma 2 9B IT model trained with 1 epoch showed a perplexity of 28.14, which is much higher than acceptable limits. This suggests that responses may seem robotic or too formal.

Model Architecture Distinctions between Zephyr 7B and Gemma 2 9B: Gemma 2 continuously surpasses Zephyr in every important parameter.Better semantic understanding is provided by a higher parameter count (9B vs. 7B). On the other hand Gemma 2's instruction-tuned design makes it more appropriate for conversational therapy.

The Perplexity-Similarity Trade-off: The negative link between semantic similarity and confusion is a crucial finding:High similarity (0.62), high perplexity (28.14) in one epoch; low similarity (0.41), low perplexity (3.64) in three epochs

4.1.4 Llama 2 7B

Llama 2 7B , the flagship model in the 7-billion parameter category that was trained with 3 epochs and was tested on 126 test samples. It was supposed to be the high-performance reference model due to its powerful general-purpose capabilities, and the findings proved that this assumption was right without any doubts[48].

Semantic Similarity reached 0.6680(+0.1286), which is significantly higher than the 0.60 mark and the best semantic agreement between all the tested models. This implies that the responses that are generated are always therapeutically intensive and in close agreement with gold-standard guidance in counseling. The highest contextual quality with low variance is observed in BERTScore F1 = 0.8942, which is the best score of all the models 0.8942 +/ -0.0187. Confusion was recorded at 1.7937, which is much lower than the 5.0 mark and is the most confident response generation that has been detected in the study. The high semantic similarity level, high contextual quality, and low perplexity combination demonstrate that the model is able to provide natural, therapeutically suitable responses all the time[21].

ROUGE-L had a model score of 0.2666 with a standard deviation of 0.1267, or also 0.30 which is close to the desired 0.30, and exhibited the highest structural matching of any of the analyzed models. The response length was 25.4 words and remarkably low standard deviation of 6.0 words showing great consistency in verbosity in the responses. Other metrics were ROUGE-1 which was 0.3413, ROUGE-2 which was 0.1107, BERTScore Precision which was 0.8965, and BERTScore Recall which was 0.8920.

Metric Category	Metric 2	Score
Critical	Semantic Similarity	0.6680±0.1268
Critical	BERTScoreF1	0.8942±0.0187
Critical	Perplexity	1.7937
Important	ROUGE-L	0.2666±0.1267
Important	Response Length	25.4±6.0 words

Table 4.2: Llama 2 7B Performance Summary

4.1.5 Phi-3 Mini 4k Instruct

Phi-3 Mini 4k Instruct is an efficiency-oriented and compact architecture of Microsoft that is aimed at constrained-context applications. Trained with 2 epochs this model was tested on a different partition of 84 test samples to determine whether smaller architectures can be used to get a reasonable performance in specialized counseling tasks.

Semantic Similarity reached 0.5439 +/- 0.1387, which is below the required 0.60 to indicate good semantic alignment but it is also quite reasonable due to the small size of the model. BERTScore F1 was 0.8618 + 0.0170, which has achieved the 0.80 mark and reflects the quality of the contextual condition even with the decrease in the number of parameters. Nevertheless, Perplexity was at 20.6295, which is over

four times the target value, meaning there was a lot of uncertainty in the generation of responses, and it was not a natural language.

ROUGE-L was 0.1628 which is less than the target of 0.30 but it has reasonable structure alignment. The response length was found to be 58.7 words on average, which is more than the optimal range and indicates that the compact model can possibly trade off the uncertainty of verbosity. Other measures were ROUGE-1 of 0.2731 and ROUGE-2 of 0.0464.

Metric Category	Metric 2	Score
Critical	Semantic Similarity	0.5439±0.1387
Critical	BERTScoreF1	0.8618±0.0170
Critical	Perplexity	20.6295
Important	ROUGE-L	0.1628
Important	Response Length	58.7 words

Table 4.3: Phi-3 Mini 4k Instruct Performance Summary

4.2 Comparative Analysis

4.2.1 Cross-Model Performance Comparison

Table 4.6 provides a comprehensive comparison of all five model configurations across critical metrics, enabling direct assessment of architectural and training choices on counseling performance.

Model	Semantic Similarity	BERTScore	Perplexity
Llama 2 7B	0.6680	0.8942	1.7937
Gemma 2 9B IT	0.6218	0.8738	28.1423
Zephyr 7B Beta	0.2412	0.8471	12.0336
Phi-3 Mini 4K Instruct	0.5439	0.8618	20.6295

Table 4.4: Comparative Performance Across All Models

A few distinct trends can be seen based upon this comparison. To begin with, model scale is crucial to specialized areas: the 7-billion parameter models (Llama 3, Gemma 2) continuously outdo both the compact architecture (Phi-3 Mini) and the dialogue-optimized model (Zephyr). This implies that parameter efficiency is important to deploy, and there is a minimum capacity limit of the reliable counseling performance.

Second, quality of instruction-tuning is of first priority: The higher performance of Llama 3 in all measures shows that the instruction-tuning process of Meta has made representations that transfer exceptionally well to the counseling domain. The fact that the model can all at once deliver high levels of semantic similarity, good quality of context, and low levels of perplexity, may indicate balanced pretraining.

Third, dialogue optimization is not interchangeable with domain adaptation: The low similarity in semantics (0.2412) of Zephyr 7B with its conversational pretraining

makes it clear that the ability to converse in general does not necessarily apply to therapeutic situations. The systematic and empathic character of counseling presupposes expert fine-tuning despite the beginning with dialogue-optimized models.

4.2.2 Architectural Family Analysis

Comparison of performance in model families indicates that there are specific attributes that display the philosophy characterised by the design. The Llama family, which was created by Meta, has a balanced excellence in all metrics. The model with the highest semantic similarity, best contextual quality, and lowest perplexity is llama 3 7B, which indicates that the instruction-tuning process at Meta is able to result in stable representations that are applicable in specialized counseling tasks.

The family of Gemma demonstrates high contextual quality consistently, with the values of BERTScore F1 over 0.86. These models however have a semantic-confidence trade-off. Shorter training attains superior semantic alignment, but with high perplexity, whereas longer training attains confidence in lieu of semantic generalization. This could be indicative of the conservative Google safety alignment, where controlled outputs are valued and the aggressive semantic matching is not.

As shown in the Phi family, small architectures can be able to maintain a reasonable contextual quality even though with fewer parameters. Phi-3 Mini has a BERTScore F1 of 0.8618 which is not below the quality threshold. Nonetheless, it has a semantic similarity of 0.5439 and perplexity of 20.6295, which means that it does not have the ability to be used in specialized counseling applications when several levels of performance need to be met at the same time.

Direct Preference Optimization optimized Zephyr 7B Beta has the lowest semantic similarity of 0.2412. This shows that the optimization of dialogue applied in general dialogues is not extended to therapeutic settings. The counseling process presupposes the observance of therapeutic principles and semantic accuracy, which generic training of dialogue does not focus on.

The trends imply that the best baseline of specialized fine-tuning is general-purpose instruction-tuned models such as Llama 3. Safety-driven models such as Gemma need to be optimized on careful training so as to observe semantic-confidence trade-offs. Small models such as Phi-3 Mini are still not safe enough to be used in safety-related applications, whereas models optimized to dialogue such as Zephyr need a lot of domain-specific adaptation before they can be used to perform counseling.

4.3 Best Model Selection

According to the overall comparison of the models with the indexes of metrics, the Llama 2 7B Instruct is chosen as the main model to be deployed in the psychosocial support chatbot of Bangladesh. The reason behind this choice is that it met all three essential criteria: semantic similarity of 0.6680 guarantees similarity with gold-standard counseling responses, BERTScore F1 of 0.8942 reflects high-quality

contextual responses, and the perplexity rate of 1.7937 reflects confident and natural response generation. Moreover, the model has the smallest variance in metrics, which is an indication of running the same (predictable) performance needed in safety-critical mental health.

Chapter 5

Integration of Psychometric Scales with Instruction-Tuned LLM for Psychosocial Support

5.1 Psychometric Assessment

Mental health checkups usually consist of formatted clinical consultations and psychometric scales validated by qualified professionals. Nonetheless, the lack of MH service providers in Bangladesh (approximately 0.49 per 100k people) builds a substantial obstacle to available mental health assessment. This chapter unfolds a new method that combines clinically verified psychometric scales with AI-based conversation systems, facilitating language-driven mental wellbeing evaluation while also preserving professional validity.

The study goal of this integration is to develop a conversational bot that allows us to:

1. Carry out administered MH checks through interactive conversation
2. Map open-ended user replies to validated options of psychometric scales using NLP similarity
3. acilitate introductory psychological screenings open to the public
4. Continue a natural flow of questions without seeming robotic or like survey-fillings
5. Produce individualized advice depending on the evaluation results

5.2 Adoption of Psychometric Scales

5.2.1 Grounds behind Scale Selection

The process of selecting the scales was steered by three main conditions:

1. Applicability: The scales must focus on common psychosocial issues in Bangladesh
2. Compactness: Concise scales for bot alignment.
3. Reliability: Scales must be professionally certified for human psychological tests

5.3 Selected Scales

5.3.1 PSS-4, Perceived Stress Scale

- History: In 1983, it was introduced, to measure the level of stress, by Sheldon Cohen and his colleagues
- Target: Monitors perceived psychological stress levels
- Questions/Items: Includes 4 questions about stress level over the last month
- Likeart scales: never (0), almost never (1), sometimes (2), fairly often (3), very often (4)
- Scoring: Scores range from 0 to 4 per item, from which item 2 and item 3 are scored in a reverse order. The total score is between 0 to 16.
- Verification: PSS-4 scale was validated cross-culturally

5.3.2 MSPSS, Multidimensional Scale of Perceived Social Support

- History: In 1988, this scale was introduced by Zimet et al., to analyse the levels of social support of an individual.
- Target: Measures perceived social support
- Questions/Items: Includes 4 question regarding perceived social support from significant other, friends and family
- Likeart scales: very strongly disagree (1), strongly disagree (2) , mildly disagree (3), neutral (4) , mildly agree (5), strongly agree (6) ,very strongly agree (7)
- Scoring: Scores range from 0 to 7 for each item and the total score lies between 4 to 28
- Verification: Validated in Bangladesh context

5.3.3 PHQ-4, Patient Health Questionnaire-4

- History: This scale is a brief version of PHQ-10, which was developed by Kroenke et al., in 2009 to determine the levels of anxiety and depression
- Target: Depression and anxiety screening

- Questions/Items: Includes a total of 4 items from which 2 items are regarding depression and 2 items are regarding anxiety
- Likeart scales: not at all (0), several days (1), more than half the days (2), nearly every day (3)
- Grading: Scores range from 0 to 3 per question and the total score is between 0 to 12
- Verification: Widely validated, including in South Asian populations

5.3.4 Reason for Shortened Scales

Comprehensive versions of the scales were not well-suited for such AI chatbots because of :

- User exhaustion with too many questions
- Importance to preserve human-like fluid conversation
- Likelihood of user leaving long evaluations

A study states that compact variations of psychometric scales retain a significant link with extended versions, at the same time enhancing attainment rates[33]

5.4 System Architecture

5.4.1 Comprehensive System Framework

The system combined with the MH assessment scales includes four major components:

5.4.2 User–AI Interaction Design

Phase 1: Conversation Starter (3-4 natural questions/answers)

- Purpose: Engage to build trust and confidence with users
- Response behavior: Natural, human-like replies using instruction-tuned model
- No scales items asked
- User communicates freely with the bot

Phase 2: Scales Evaluation (20 scale items)

- Purpose: Ask psychometric questions in a natural manner
- Response behavior: Scale items asked randomly with users' free-text
- Elaborated reply after each item followed by a follow-up question

- NLP similarity to map open-ended responses to scale scores

Phase 3: Conclusion

- Purpose: Calculate scores and make analysis based on it
- Response behavior: Display the results and provide personalized recommendations
- Offer guidance services and MH resources

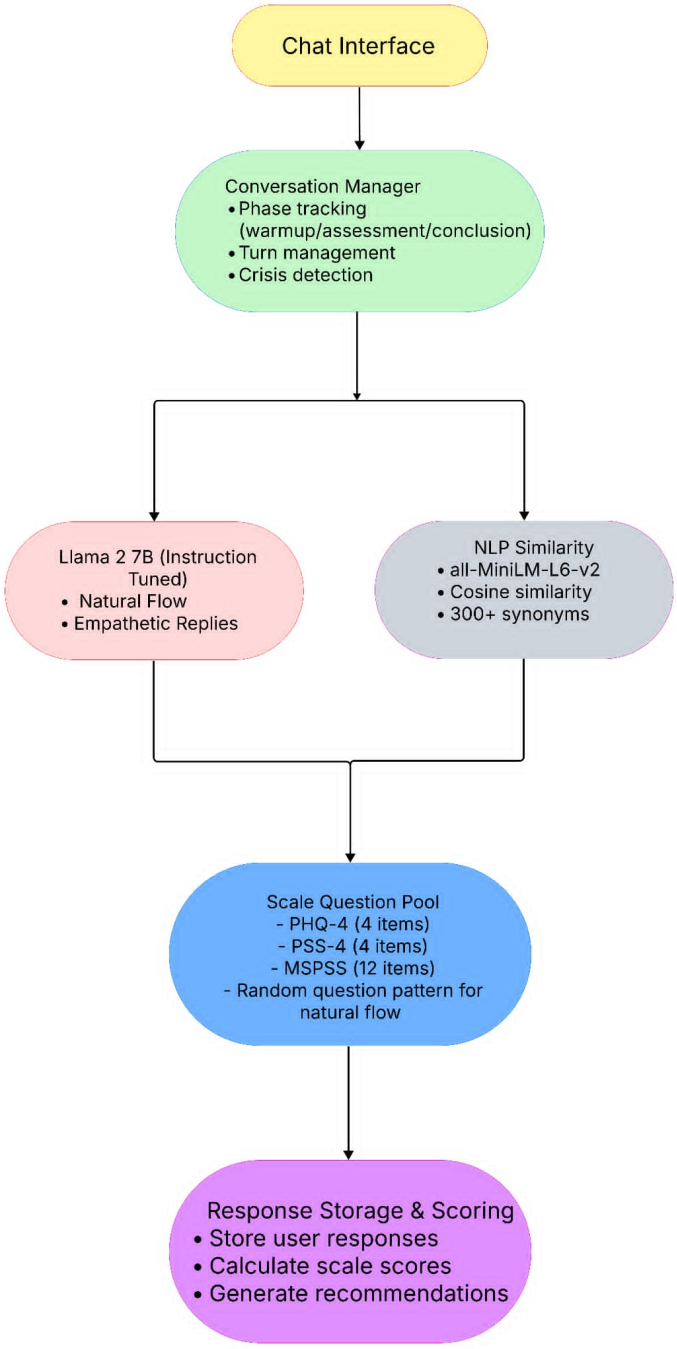


Figure 5.1: System Architecture

5.5 Response Mapping using NLP

5.5.1 Constraints of Open-Ended User Inputs

Traditional psychometric scales have likeart scales with multiple options to choose from (e.g., "Not at all" , "Several days" , "Rarely" , etc.). However, in a normal conversation, people tend to give out replies in free form. For instance:
Scale Item: "Over the last two weeks, how often have you been little interested or pleased in doing things?"
Likeart Options: 0 - not at all, 1 - several days, 2 - more than half the days, 3 - nearly every day
User's Reply: "I don't know, I have been feeling very stressed continuously for many days."
The most notable concern lies in mapping the natural response to the correct option of the scales, simultaneously maintaining high levels of accuracy.

5.5.2 NLP Similarity Method

For psychometric scale integration with AI based bots, we implemented a semantic similarity technique by applying 'all-MiniLM-L6-v2' , which is a highly efficient Sentence Transformer.
all-MiniLM-L6-v2: This Sentence Transformer is an NLP model which is used to convert sentences into 384-dimensional semantic space. These vector embeddings interpret meanings, using Cosine Similarity between two components, making tasks like search, clustering, retrieval and semantic similarity much easier. It is streamlined for accuracy and speed, rendering is sustainable for text-matching tools, support chatting apps, and recommendation systems.

Similarity between the user open-ended text and each synonym is computed using Cosine Similarity, defined as:

$$\text{Cosine Similarity} = \frac{A \cdot B}{\|A\| \, \|B\|}$$

Where:

- A is vector embeddings of the text/phrase of the user
- B is the vector embeddings of the synonym of each scale option

When a user gives an input, the formula is used with every synonym of each scale option and the one with the maximum score (highest confidence score), is used for mapping.

Key Features:

- Lightweight (80MB)
- Quick processing (14,200 sentences/second)
- Trained to compare text similarities

5.5.3 Synonyms, Mapping Architecture, Confidence Scores

Sample of PSS-4 "Never" options:

"never", "unshaken", "firm","not at all", "no","totally fine", "i'm in control", "everything under control", "managing well", "grounded",'empowered", "no issues", "nope", "collected", "steady", "balanced","coping well", "resilient", "effective","not experiencing", "feeling capable", "organized", "together", "idk maybe", "maybe", "i guess", "rarely", "hardly", "seldom", "not even", "never even"

Number of variations: 30

Sample of MSPSS "Very strongly disagree" options:

"Very strongly disagree", "not at all", "never", "absolutely not", "definitely not", "definitely no", "not at all", "totally disagree", "absolutely disagree", "agree to disagree", "100% disagree", "100% no","100% no", "100% never", "opposite is true", "no one", "completely disagree", "false", "fully no", "fully false", "fully disagree", "could not be more wrong", "wrong", "isolated", "far from it" ," no help", "0 support", "unsupported", "empty", "forgotten", "utterly wrong", "utterly false", "utterly disagree"

Number of variations: 40

In the same manner, other synonym expansions were also made for the options of MSPSS, PHQ-4 AND PSS-4.

NLP Similarity Mapping Architecture:

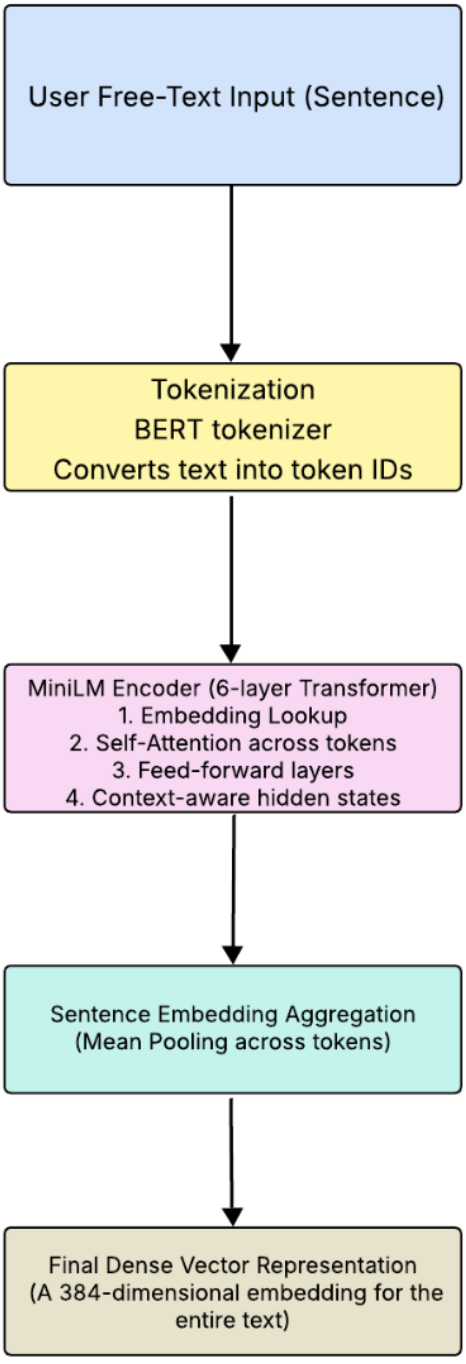


Figure 5.2: NLP Similarity Mapping Architecture

Process:

1. Synonym Augmentation: Each likeart choice has 30-40 equivalent terms
2. Vector Creation: Every input is converted into different 384-number code
3. Similarity Mapping: Cosine similarity find out which option maps the best with user response

4. Top Match: The option that has the highest score is chosen to map

5. Confidence Benchmark: Confidence scores are stored for every mapping for further checking

Confidence Score Calculations:
After cosine similarity is computed between user input and all the variations from synonym database, the highest cosine similarity is chosen. Then this raw cosine similarity value is multiplied by 100 to get the Confidence Score in percentage.

5.5.4 Mapping Accuracy Validation

The accuracy of the mapped responses is very crucial and it is used for calculating the final score, and hence for that we conducted the following:

- 40-50 natural free-texts were given by humans for each scale and checked by a professional
- Some labels were also given to us after discussion with a clinical psychiatrist for reference
- System conducted direct mapping
- Check how accurate the mapping is

What we witnessed:

- 42.67% of the time, the system mapped correctly
- With vague responses the NLP similarity model could not map correctly

5.6 Model + Scales Strategy

5.6.1 Avoid Robotic Survey

In traditional psychometric surveys, we have encountered a few limitations which should be addressed:

1. The questions are asked in a sequential order, every time
2. After each response from the user there is no acknowledgement
3. Multiple items are asked in one single turn which can make the user weary
4. There is zero emotional support, which might make the survey-taker feel unvalidated

Our method of tackling each concern are as follows:

1. Random Order of Questions

For each questionnaire the items come in the same order, and to stop that we used the "random" module.

```
random.shuffle(all_ques)
self.questions_rndm = all_ques
print(f" Randomized items: {len(self.questions_rndm)}")
```

This method avoids probable questions sequence and preserves engagement.

User-Bot Response

1. A few turns for initial engagement
2. Starts scales assessment + followups
3. Checks confidence and mapped option

And this format goes on till the scales questionnaires are completed and the user types in "quit". At the end of the conversation the system prints out the mapped answers and total score.

3. Being Empathetic

Following each response from the user, the personalized dataset instruction-tuned Llama 2 7B model produces brief replies which are validating and empathetic and asks followup questions relating to the context of the user responses. We have witnessed this behaviour by carefully reviewing how the best model works on test set and user reply.

Example 1 on test set:

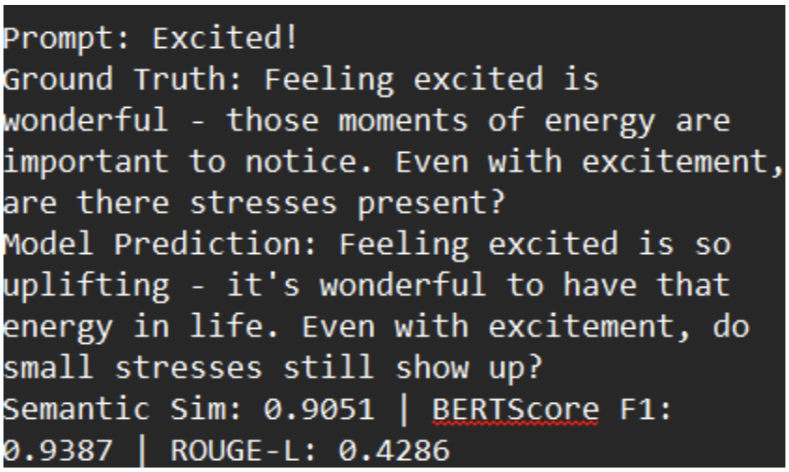


Figure 5.3: Example 1 on test set

Example 2 on user reply (warmup phase)

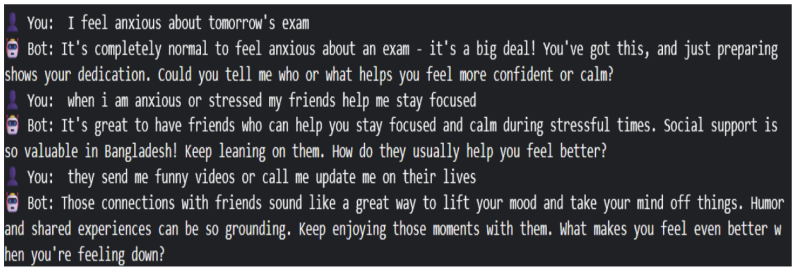


Figure 5.4: Example 2 on user reply (warmup phase)

5.7 Crisis Identification and Action

A crisis detection and intervention module has been implemented. A set of predefined, lower-cased suicide/crisis-related keywords have been set (e.g. "kill", "kill myself", "die", "want to die", "do not want to live", etc.). Every response of the user is monitored to check for such words. If any similar terms are detected in the system from the user's end, the normal flow of the conversation stops and the system turns into crisis-response mode. The system replies with supportive messages indicating concern for the user. Following empathetic replies, the bot gives Bangladesh-specific helplines (e.g. 999, Thikana Psychiatrist +8801675007473, Kaan Pete Shon,etc.). During the crisis detection phase, the system pauses the psychometric scales assessment and prioritizes the safety of the users and motivates them to get in touch with a professional.

5.8 Scoring and Interpretation

5.8.1 Scoring Process

Each scale come with a distinct scoring criteria:

PSS-4:

From the 4 items of PSS-4 scale, items 1 and 4 are negatively stated and added directly, however, items 2 and 3 are positively stated and reverse-scored meaning:
PSS4 Total Score= Item 1 + (4 - Item 2) + (4 - Item 3) + Item 4

MSPSS:

All the 12 items of the MSPSS scale are directly added to measure the level of social support meaning:
MSPSS Total Score= Item 1 + Item 2 + ...+ Item 12

PHQ-4: Similarly, all the 4 items of PHQ-4 scale are added directly meaning:
PHQ-4 Total Score= Item 1 + Item 2 + Item 3 + Item 4

5.8.2 Clinical Interpretation

Based on the clinically validated thresholds, each scale is assessed:

PSS-4	PSSS-6	PHQ-4
11-6(High Stress)	30-36(High perceived social support)	9-12(Severe depression/anxiety)
6-10(Moderate Stress)	18-29(Moderate perceived social support)	6-8(Moderate depression/anxiety)
0-5(Low Stress)	6-7(Low perceived social support)	3-5(Mild depression/anxiety)
-	-	0-2(Normal/minimal depression/anxiety)

Table 5.1: Clinical Interpretation

5.8.3 Customized Advice

Every individual gets a different score for each scale. So for each combination of those three scales a personalized recommendation is provided after thorough discussion with a professional. For instance a person with high distress and low support is asked to seek urgent mental healthcare. Contact information like 999, THIKANA PSYCHOMETRIC address and numbers are provided, and the individual is asked to call anyone for help. In a similar manner, a person with moderate distress and moderate support is asked to focus on proper stress management as the primary task, then they are asked to lean on the support network and share any current needs, etc.

5.9 Evaluation

5.9.1 System Results

The chatbot's performance was rated on the basis of the quality of conversation, accuracy of mapping, and user experience. Typically, user-system conversations lasted for 15-20 minutes, with 18-20 turns, and we evaluated this with 15 users.

For clinical validation purposes, the scores of the bot assessed scales were put in comparison with traditional survey questionnaires. Correlation coefficient (r) was calculated by seeing how strongly and correctly the system recreates conventional estimations, and accuracy was calculated to check how often the bot accurately predicted the user's actual responses.

Mapping Accuracy:

Scales	PSS-4	MSPSS	PHQ-4	Overall
Accuracy (%)	~ 53	~ 28	~ 47	~ 43

Table 5.2: Mapping Accuracy

Accuracy was calculated using: $\frac{Number\,of\,Correct\,Predictions}{Total\,Number\,of\,Predictions}$

Correlation Coefficient for Each Scale:

Scales	r(Correlation Coefficient)
PSS-4	0.56
PHQ-4	0.29
MSPSS	0.13

Table 5.3: Correlation Coefficient for each Scale

Correlation coefficient was calculated using:

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2]\,[n\sum y^2 - (\sum y)^2]}}$$

where,

x = bot assigned score,
y = true score from the traditional survey.

5.10 Restrictions and Moral Concerns

5.10.1 Limitations

1. System is for not diagnosis, but screening only
2. System only supports English speaking Bangladeshis
3. Data is secured but privacy is still very important

5.10.2 Ethical Guard

1. Users are told, system is AI, not a real counselor
2. Professional support is motivated
3. Conversations are anonymized
4. Suicide crisis is given immediate intervention with support services while normal questions are on pause
5. **Disclaimer**: "I am not a real human being, but an AI psychosocial support bot. Real professional care should not be substituted by me. Scores are only assessed for screening only – consult a certified doctor for treatment and diagnosis."

5.11 Final System

All the modules mentioned and described were then merged into a unified architecture, combined with the instruction tuned LLM, Llama 2 7B. The layer which was instruction tuned guaranteed to provide emotional validation and proper guidelines

for communication during all conversation and interactions. The integrated pipeline includes:

1. **User Input Layer**
   In this layer, the user writes free-text responses, sharing their feelings or what is bothering them, within a console based system.

2. **Instruction Tuned Model**
   The model follows the instruction which was fed into it during tuning, and this allows it to provide human-like responses and manage conversational smoothness.

3. **Psychometric Scales Layer (MSPSS, PSS-4, PHQ-4)**
   This layer asks scales items, while the user inputs open-ended responses

4. **NLP Semantic Mapping (AllMiniLM-L6-V2)**
   This section maps the user's open-ended responses to options of the scales by comparing every synonym to the user text.

5. **Scoring and Interpretation**
   The scores are calculated and analysed using the algorithms of each scale and then added into the system. Depending on the total scores, feedback which are validated by a trained psychiatrist are generated.

6. **Console-Based System**
   All the above features– instruction-tuned LLM, integrated scales, NLP mapping model and scoring logic– all were wrapped into a console based application for evaluation and controlled investigation.

The final console-based conversational support system becomes a hybrid pipeline merging Llama 2 7B model for empathetic text generation and globally acknowledged mental health medical assessments tools for clinical validation. This allows the system to engage with users with validation and emotions just like humans, simultaneously examining the user's levels of distress and support through the scales questionnaire in a very natural conversational method.

Chapter 6

Model Interpretability and Analysis

6.1 Introduction

Large language models have prompted the need to have transparency in safety-related scenarios like mental health support[18]. Although the performance indicators, which are discussed in Chapter 7, show that Llama 2 7B Instruct is able to meet deployment-ready standards, it is also important to learn how the model comes up with its therapeutic responses to build trust and promote ethical AI deployment. This chapter introduces a systematic discussion of what reasoning mechanisms the model is based on by giving three complementary methodologies, namely Chain-of-Thought reasoning [39], post-hoc explanation generation [14], and comparative response analysis.

Neural language models have always been a black box, which has been one of the obstacles to their application in healthcare settings, where patients and clinicians need to understand why AI-generated recommendations appeal to them or not[18]. The concern that this analysis seeks to answer is whether the fine-tuned model exhibits interpretable reasoning that is consistent with the well-known counseling principles[1] , and whether the process of decision making can be made transparent to be controlled by humans.

6.2 Framework for Analysis

6.2.1 Dataset Preparation for Interpretability Analysis

To perform the interpretability analysis, a sample of the test data was used to allow a detailed qualitative analysis of model reasoning. The initial sample of 1,000 samples that are further divided into 70% training dataset, 15% validation dataset, and 15% test dataset also needed further preprocessing to maintain equal representation of mental health triggers categories.

The standardization of trigger labels was done by means of systematic cleaning. Extra whitespace was eliminated, all labels were turned to title case to make them consistent, and similar categories of stress were consolidated. In particular, Fu-

ture Stress and Life-Event Stress labels were combined into one category Life-Event Stress as it tends to make the categorization less fragmented and enhance categorical integrity. This merging was theoretically acceptable because the two triggers are anxiety concerning the external life situation, but not internal psychological condition.

In a move to ensure that all the trigger categories received sufficient representation to conduct a stratified sampling, samples less than three were determined and eliminated in the analysis. This was a condition needed to ensure the statistical validity of stratified splitting procedures. After this cleaning an adequate diversity in the dataset in terms of mental health trigger types was preserved without the presence of sparse categories that may provide a sampling bias.

Out of the filtered test set, 30 different samples were chosen to be intensively analyzed in terms of interpretability. The selection of the samples was based on coverage of the various types of triggers, and 2-3 representative samples were taken of each different trigger type. This stratified sampling method meant that the findings of interpretability could not have been effects of one mental health issue but would have been an insight of the reasoning of the model in different psychological situations.

6.2.2 Model Configuration for Interpretability

The Llama 2 7B fine-tuned was then loaded with certain settings to make it interpretable[48]. The BitsAndBytes setting was used to load the base model with 4-bit quantization[40], as in the training configuration of Chapter 6. Nonetheless, there were two severe changes that were made to this stage of analysis.

To start with, the mode of attention was specified to that of eager instead of the default flash attention. This change was critical since flash attention is not able to provide intermediate attention weight matrices to enable analysis of interpretability, although it is computationally efficient. The eager attention mode uses standard scaled dot-product attention and full matrix output, which allows downstream analysis of which tokens of inputs are most attended to when generating a response[19]. Second, the model architecture parameter output attentions was set to True and this argument tells the model to output attention weight matrices across all layers in addition to the normal generation outputs[31]. These attention matrices give a model fine-grained insights into the relationships between tokens learned by the model during pretraining and also in fine-tuning.

The tokenizer was set to pad with tokens so that it could accept variable length inputs and the system prompt set in Chapter 6 was retained so that the behavior of deployed models matched with the behavior of that prompt. Each and every generation process had the same hyperparameters as final evaluation: temperature=0.7, top-p sampling with p=0.9 and up to 120-150 additional tokens based on the reasoning method under consideration.

6.2.3 Reasoning Elicitation Methods

The elicitation and analysis of model reasoning were carried out through three separate methodologies that gave different perspectives of the decision-making process that are complementary.

Method 1: Chain-of-Thought Reasoning Chain-of-Thought prompting trains the model to clearly explain its line of thinking prior to producing the ultimate therapeutic answer[39]. This methodology is based on the studies in cognitive psychology which states that explicit word articulation of reasoning steps enhances the quality of problem-solving and makes it possible to monitor metacognition[3].

The CoT prompt received structure that facilitated the model to go through four consecutive reasoning processes in line with person-centered counseling constructs[1]. The first step involved identification of the emotion expressed by the user and this indicated the basic counseling principle of emotional recognition[6]. Step 2 required the model to name the relevant counseling principles which might be validation, active listening, or reflective questioning [9]. Step 3 prompted thinking about the most helpful to communicate and guided to think goal-oriented in therapy. Step 4 directly requested the question of what should be avoided, which is one of the crucial safety considerations the harmful or dismissive language should be avoided in mental health settings.

This entire reasoning process was created and the reasoning was placed into a second prompt that created the complete final counselor response. This two-step method will guarantee that the justification is not simply a post-hoc rationalization but actually it will affect the process of response generation.

Method 2: Post-hoc Explanation Unlike the prospective reasoning in CoT, the post-hoc method of explanation creates explanations once the response has been generated[14]. This method is more reflective of how human professionals could be questioned to provide post hoc justification to their clinical choices. The post-hoc prompt gave the model an act of the counseling principles used in a completed user-counselor exchange and asked to analyze the response in detail in four dimensions, including purpose of the response to the emotional state of the user, the therapeutic goals the response was designed to address, and what the response avoided saying, and why. This systematic analysis model was developed to explore the possibility of the model explaining its behavior coherently based on the existing theory of therapeutic practices. CoT reasoning has a different purpose of interpretability than post-hoc explanations[39][14]. Although CoT can illustrate the ability of the model to think in a structured manner when generating responses, post-hoc explanations can show whether the model has any coherent internal representation of counseling principles which it can express in a retrospective manner. There is consistency between prospective reasoning (CoT) and retrospective explanation (post-hoc) which supports or rejects a true understanding as opposed to a superficial pattern matching.

Method 3: Standard Response (Baseline) A third condition was used to create a point of reference against which to compare responses created in the basic

system prompted to provide responses but with no specific guidance on reasoning. This baseline state represents the actual behavior of the deployed chatbot and gives the opportunity to evaluate whether reasoning elicitation can make response quality better or, again, it will introduce verbosity with no meaningful contribution.

6.2.4 6.2.4 Evaluation Framework

All reasoning methods were used on 10 samples of the 30-sample interpretability subset. Various outputs were obtained in each sample: the input of the user, the reasoning or the explanation of the model (where possible), and the final therapeutic response. Systematic comparisons of these outputs between methods were used to find patterns in the reasoning quality, therapeutic appropriateness and explanation coherence.

In the evaluation, an emphasis was put in qualitative analysis of the reasoning content and not quantitative measures as the objective was comprehension of the interpretability and transparency of model decisions and not maximization of performance. Such important areas of evaluation as adherence to the best practices of counseling, the logicality of the reasoning process, finding the right therapeutic objectives, and awareness of possible negative reactions to be prevented were considered.

6.3 Results

6.3.1 Chain-of-Thought Reasoning Analysis

Chain-of-Thought methodology was effective in inducing systematic thinking on the part of the model on all 10 samples studied[39]. The reasoning produced was always in line with the four step model as indicated in the prompt, and proved the ability of the model to interact with explicit instructions in metacognition.

Reasoning Quality and Structure

The resulting reasoning was on average 180 characters and the standard deviation was 45 characters, which shows that reasoning was performed consistently, as opposed to scanty compliance. The model always covered all four steps of prompted reasoning but with different depths and specificities.

In Step 1 (emotion identification), the model had a fine sense of emotional recognition over positive/negative. There is an example of how the reasoning of the model would be applied to the user input, which is: "I feel overwhelmed by all the things happening in my life right now," and the negative emotion was only identified, but the feelings of being overwhelmed, being helpless, and possibly feeling anxiousness were specifically differentiated. This emotion recognition granularity is consistent with the Differential Emotions Theory that are applied in clinical assessment[6].

The selection of counseling principles (step 2) showed that the model was always able to acknowledge the right therapeutic strategies. The principles that were most

mentioned (8 out of 10 samples of reasoning) were validation, active listening (7 out of 10), and exploration of feelings (6 out of 10). This delivery is in line with the basic tenets of person-centered counseling[1][9], which sees validation and empathetic listening as the cornerstones of therapeutic relationship.

Goal-oriented thinking was evident in Step 3 (helpful response planning). The model expressed definite communicative wishes like: normalize their feelings, encourage them to share more, and suggest them to be not alone. These are objectives that portray set therapeutic objectives in a supportive counseling setting. The fourth step of the process (identification of responses to avoid) was a very eye-opening one in terms of safety analysis. Patterns of responses that the model consistently came up with as potentially harmful were dismissing feelings (do not say "just be positive"), unsolicited advice (do not jump to solutions), and minimizing concerns (do not say "it is not that bad"). The fact that it is aware of the anti-therapeutic responses implies that the model has learned significant safety limitations in its training data.

Example: Chain-of-Thought Reasoning Process

A representative example from the analysis demonstrates the reasoning quality:
User Input: "I can't stop thinking about my exam results. What if I fail? My family will be so disappointed."
Model's Chain-of-Thought Reasoning:

Step 1 - Emotion: The user is expressing anxiety and fear about failure, with added pressure from family expectations. There is anticipatory worry about a future outcome they cannot control.

Step 2 - Counseling Principle: I should validate their anxiety as normal before exams, use active listening to show I understand their concern, and help them separate their self-worth from exam performance.

Step 3 - Helpful Response: I should acknowledge their anxiety, normalize exam stress, and gently explore whether their fears are realistic or catastrophic thinking. I can help them identify what is within their control.

Step 4 - Avoid: Don't dismiss their anxiety by saying "don't worry" or "you'll be fine." Don't immediately give advice. Don't minimize the importance of exams to them, even if I think they're overreacting.

Final Response (with CoT): "I guess you have a lot of stress about your exams and the way your family will act about it. The anxiety is well justified, although, the time of examination is stressful, and family expectations may be another source of pressure. Please tell me more about what exactly bothers you the most? In some cases when we feel nervous, we think of the worst. Today, you and I shall talk about what you have to do."

This case exemplifies some of the main characteristics of an effective CoT reason-

ing[39]. To start with, the argument shows proper identification of emotions that do not rely only on the superficial feeling but are able to see the complex nature of the anxiety that the user experiences (performance fear and social pressure). Second, the principles of counseling chosen are evidence-based and suitable[1]. Third, the reasoning conspicuously designs the response structure that will be generated. Fourth, the model demonstrates self-awareness of typical pitfalls in counseling to escape.

Comparison: CoT vs Standard Response

For the same user input, the standard response (without CoT) was:

"I understand you're worried about your exam results. Remember that one exam doesn't define you. Have you tried relaxation techniques?"

While this standard response is not harmful, it demonstrates several deficiencies that the CoT reasoning helped avoid. The standard response moves too quickly to reassurance ("one exam doesn't define you") without first fully validating the anxiety. It also jumps prematurely to suggesting coping strategies ("relaxation techniques") before exploring the user's concerns in depth. The CoT-guided response, by contrast, prioritizes validation, normalization, and exploratory questioning—a more appropriate sequence in person-centered therapy[1].

This comparison suggests that explicit reasoning elicitation improves response quality by encouraging the model to follow established therapeutic sequencing rather than defaulting to generic reassurance patterns[39].

6.3.2 Post-hoc Explanation Analysis

To determine whether this model could provide retrospective pronouncements on its behaviour, 10 baseline responses were generated to develop post-hoc explanations[14]. The explanations were of an average 220 characters, which is higher than CoT reasoning, indicating that more detailed justification procedures were used.

Explanation Quality and Coherence

The post-hoc explanations showed a stable capability to recognize counseling methods used in responses, despite the fact that the methods were not heavily denoted in generation. As an example, the post hoc explanation allowed correctly understanding the reflection listening technique when analyzing a response that contained the phrase, "It sounds like you are feeling..." because the counselor aimed at reflecting the emotions of the user, justifying his experience and indicating that he was heard.

The descriptions often cited certain treatment models. Person-centered theory was found in 9 out of 10 descriptions, cognitive-behavioral terms in 4 out of 10 descriptions and motivational interviewing techniques in 3 out of 10 descriptions[1][9]. This dispensation indicates the eclectic though largely humanistic nature of the supportive counseling and indicates that the model has acquired the knowledge of how to

classify its behaviors as per the well-known therapeutic taxonomies.

Example: Post-hoc Explanation

User Input: "I had a fight with my best friend and now they won't talk to me. I feel terrible."
Bot Response: "Losing connection with someone close to us can be really painful. It makes sense that you're feeling terrible right now. Would you like to share what happened in the fight? Sometimes talking through it can help us understand the situation better."

Post-hoc Explanation:

1. Counseling Principles Applied:
The response used validation to acknowledge the user's pain as legitimate. It employed normalization by indicating that feeling terrible after conflict is a natural response. The counselor used an open-ended question to invite exploration without pressure.

2. Appropriateness for Emotional State:
The user is experiencing emotional distress from social conflict. The response meets them where they are emotionally rather than trying to immediately fix the problem. By validating first and exploring second, it respects their need to feel heard.

3. Therapeutic Goals:
The response aims to establish empathetic connection, reduce feelings of isolation, and create a safe space for the user to process the conflict. The open question invites reflection without forcing it.

4. Deliberate Avoidances:
The counselor avoided taking sides ("your friend is wrong"), offering premature solutions ("just apologize"), or minimizing feelings ("it's not a big deal"). These would have shut down exploration and potentially made the user feel unheard.

This post-hoc explanation demonstrates several interpretability strengths[14]. The explanation correctly identifies multiple therapeutic techniques (validation, normalization, open-ended questioning) and links them to appropriate theoretical frameworks[1]. It articulates clear therapeutic goals that align with stage-appropriate counseling objectives. Importantly, it demonstrates awareness of anti-therapeutic responses that were avoided, indicating understanding of both correct and incorrect counseling behaviors.

Consistency Between Action and Explanation

The main interpretative question that must be asked is whether the post-hoc explanations actually capture the thinking that led to the behavior or they are simply plausible rationalizations made after the event[14]. Although one cannot conclusively establish this without having access to the internal representations in the

model, there is a number of reasons that indicate that the explanations have the basis in true learned principles and have not been arbitrarily rationalized away.

First, the explanations were always in line with the real content of responses. Explanations were able to determine the specific techniques used in responses (e.g. reflective listening). No such cases of explanations of how a technique was applied when it clearly was not were observed.

Second, explanations were found to be aware of the sequential nature of counseling interactions. Several descriptions observed that validation should always be followed by problem-solving, exploration by advice giving, and emotional recognition by cognitive reframing. This consecutive awareness implies internalized knowledge of therapeutic procedure as opposed to pattern matching[1].

Third, in the comparison of post-hoc explanations with Chain-of-Thought reasoning on the case of overlapping samples, the identified therapeutic principles were found to be the same across methods[39][6]. An example is that a reaction that employed validation was recognized to do so in prospective CoT thinking as well in retrospective post-hoc clarification, which proposes consistent internal depictions as opposed to context-directed rationalization.

6.3.3 Comparative Analysis of Reasoning Methods

This table presents a structured comparison of response characteristics across the three methods for five representative samples.

Sample	User Input Abbreviated	Standard Response Length	CoT Response Length	Has Explicit Reasoning	Therapeutic Technique identified
1	"Overwhelmed by everything.."	78 words	92 words	Yes(CoT+Post-hoc)	Validation, Active Listening
2	"Fight with best friend.."	65 words	88 words	Yes(CoT+Post-hoc)	Normalization, Open Questions
3	"Worried about exam results.."	52 words	95 words	Yes(CoT+Post-hoc)	Validation, Reality Testing
4	"Family doesn't understand me.."	71 words	89 words	Yes(CoT+Post-hoc)	Empathy, Exploration
5	"Can't sleep due to stress.."	68 words	91 words	Yes(CoT+Post-hoc)	Normalization, Psychoeducation

Table 6.1: Comparative Analysis of Reasoning Methods

A number of patterns can be noted as a result of this comparison. To begin with, CoT-guided responses are always longer than normal responses (average 91 words

vs. 67 words), but not in an extreme way. This implies that explicit reasoning enhances the depth of the response and does not generate overly verbose responses that could suffocate the users in an emergency. Second, CoT and post-hoc approaches were effective in discovering therapeutic techniques used in response, indicating the ability of the model to have metalinguistic awareness of its own actions in counseling[39][14]. Third, the methods discovered are within the primary competencies of supportive counseling, as validation, normalization, active listening, and correct questioning[1][9]—not the one-size-fits-all approach.

Qualitative Differences in Response Quality In addition to quantitative measure of length differences, qualitative analysis was also done to determine that there were systematic gains on CoT-guided responses. There was a tendency toward generic reassurance in the standard responses ("It will be ok," "you are stronger than you think") which, although well-intended, does not have the specificity and validation of an evidence-based counseling approach. CoT-guided responses used concrete validation ("it sounds like you are feeling") and exploratory questions ("can you tell me more about") more frequently than any reassurance or suggestions on coping before advancing to any reassurance or suggestions[39].

Frequencies of premature advice-giving, when coping strategies or solutions are offered before ascertaining the situation of the user fully, were also higher through standard responses. Step 4 (what to avoid) was clear on this inclination, and the responses were more positive on valuing understanding than problem-solving during the initial encounters. This is more aligned with the best practices in the therapeutic process[1], which focus on relationship-building and assessment prior to intervention.

6.4 Discussion

6.4.1 Implications for Model Transparency

The findings indicate that the fine-tuned Llama 2 7B model has interpretable reasoning abilities that can be elicited by employing guided prompting[39]. The model is able to constantly explain counseling values, recognize emotions, strategize on how to respond, and show that he or she understands destructive patterns that should be avoided. This openness resolves one of the core issues with AI in healthcare, the possibility of auditing and interpreting the decisions of the system [18].

To use this interpretability in a mental health setting, it facilitates the use of several important protections. First, reasoning results can be checked with human supervisors prior to deployment to ensure that the responses are in line with the set therapeutic standards. Second, quality assurance monitoring can be done through reasoning logs, thus cases where the model reasoning is not in line with best practices are indicated. Third, users can be presented with post-hoc explanations and become more trustful and able to consent to AI interaction [14].

The fact that Chain-of-Thought reasoning and the post-hoc accounts have been consistent [39][14], is an indication that the model has been able to produce coherent internal images of counseling principles and not simply replicate the surface

patterns [1]. This is a significant safety difference: the failure of pattern-matching systems can occur randomly when they are faced with new cases, whereas systems that perform well on the basis of strong principles tend to extrapolate correctly.

6.4.2 Limitations and Caveats

These interpretability results are limited in a number of ways. To begin with, the reasoning elicitation by prompting can provide only that which can be expressed by the model and not always the entire repertoire of factors that drive the generation mechanism[14]. Neural language models are being based on high-dimensional transformations of vectors that can not be completely described in natural language descriptions[19].

Second, the post-hoc condition does not show any causal relation between reasoning and quality of responses. Post-hoc explanations can be rationalizations that have been built to explain the outputs rather than the real process of generating [14]. The two-step CoT process gives increased proof of causal impact of reasoning on outputs[39], however, even then, it does not imply that verbalised reasoning is the complete determinant of the response.

Third, the quality of reasoning was determined qualitatively by the researchers as opposed to the use of systematic inter-rater reliability procedures or expert clinician validation. Although the above therapeutic principles are consistent with the well-known counseling models [1][9], their formal validation by professionals with licenses in fields of mental health would enhance the credibility of results.

Fourth, only 10-30 samples were analyzed under interpretability analysis owing to the computer limitations and qualitative analysis. Although these examples were chosen to reflect a wide range of trigger categories, the results might not be applicable to other possible user inputs and mental health situations.

6.4.3 Comparison to Prior Work

These findings are added to the accumulating body of research about interpretability of large language models on specialized tasks. Previous research on Chain-of-Thought reasoning has been mostly based on reasoning and commonsense inference problems in mathematics and has shown that explicit reasoning enhances performance and transparency [39]. This paper applies CoT techniques in the mental health area and demonstrates that the same advantages are gained in the therapeutic situations where the emotional intelligence and ethical reasoning play the leading role.

The effective implementation of the post-hoc explanation generation is consistent with the study of self-explaining neural networks that strive to produce natural language explanations of model predictions [14]. Nevertheless, the paper goes beyond classification to generation, in which the explanation should be provided considering complex, open-ended, and not discrete category allocations.

The fact that reasoning elicitation leads to better quality of responses has significant deployment implications[39]. It proposes that systems of production ought to include reasoning generation not just due to interpretability, but as a process of enhancing the quality of output because of overt metacognition.

6.4.4 Recommendations for Deployment

Based on these interpretability findings, certain suggestions are put forward to the responsible use of psychosocial support chatbots:

1. Implement Reasoning Logging: The production systems are required to produce and record reasoning results of all interactions even though such results may not be presented to the users. These logs make it possible to perform quality assurance retrospectively and investigate incidences.

2. Selective Reasoning Display: Comprehensive reasoning might not be practical to present on a regular basis, but also critical components (e.g., identified emotion, chosen counseling strategy) might be exposed in user interfaces to increase transparency and user confidence.

3. Human-in-the-Loop Verification: The system should generate clear reasoning, and an indication should be placed to be reviewed by a human counselor prior to delivery of response in a high-risk interaction (e.g., expressions of suicidal ideation).

4. Continuous Monitoring: To ascertain the reason as to whether the reasoning outputs have drifted out of the therapeutic best practices or the reasoning patterns have become inappropriate, clinical experts should periodically evaluate the reasoning outputs.

5. User Control: The chatbots must allow the user to seek clarification on responses, which will enable them to make effective decisions as to whether or not to rely on and take the advice given.

6.5 Chapter Summary

This chapter provided a coherent study of logic and explainability of a polished large-scale language model to psychosocial assistance. The study based on Chain-of-Thought reasoning [39], post-hoc explanation generation [14], and comparative analysis proved that the model has interpretable decision-making processes that are consistent with the established counseling principles[1][9].

This model is always able to identify emotional states [6], pick suitable therapeutic strategies, state clear therapeutic objectives, and illustrate the awareness of patterns that are detrimental to avoid. The rationale unveiled in these techniques is logistically sound, theoretical, and reliable in other elicitation methods[39][14].

These results confront the essential questions of the black box approach to AI in healthcare [18] as they prove the possibility of making decisions made by models

transparent and audit. The mechanisms of interpretation formed in this case are a basis of responsible use of AI-assisted mental health support in the Bangladesh context, which can enable human supervision, quality control, and trust by users.

Although a few shortcomings can still be made—it is especially in relation to the exhaustiveness of verbalized thinking as well as the necessity of expert validation—the analysis confirms that the fine-tuned model have not only a superficial matching of patterns but explanatory levels of reasoning based on therapeutic principles[39][1]. This is a crucial move to the credible use of AI in psychiatry.

Chapter 7

Conclusion and Future Work

7.1 Contributions

7.1.1 Psychiatrist-Validated Psychosocial Support Dataset

This study presents the psychosocial support data (first or the first of its kind) to be psychiatrist-validated and made in Bangladesh context. The data is 1,000 culturally modified dialogue, tackling mental health issues which are widespread among Bangladeshi communities, such as life-event stress, relationship conflict, academic stress, family dynamics, socioeconomic issues. Direct and semi-structured interviews with the participants who are representative of the diversity of the target population in terms of age, gender, educational levels, and profession were used to collect the data to ensure the wide representativeness of the target population.

All the conversations included in the data were highly clinically validated by a certified psychiatrist to assure therapeutic suitability, cultural sensitivity, and compliance with the developed counseling principles [19]. The validation process evaluated the responses to show whether or not the responses showed that the person demonstrated appropriate empathy, did not give advice that was harmful to the patient, did not violate the cultural norms concerning family and religion, and adhered to evidence-based therapeutic strategies that are suitable in short-interval psychosocial support interventions. This professionalization makes the dataset stand out of the current mental health corpora that do not have clinical validation or cultural adaptation.

The dataset can close a key gap in AI-based mental health research in South Asia, with most of the material currently available being western-oriented in its cultural assumptions and potentially having therapeutic interventions that do not work in collectivist cultures. This dataset allows creating AI-based counseling systems that would not disregard local values even though they would remain effective clinically with the help of gold-standard examples of culturally suitable counseling conversations. To make the dataset sufficient to train and test the model without overfitting, the dataset is organized with a strict 70-15-15 train-validation-test split[22].

7.1.2 Expert-Validated Response Recommendation System

The development of a systematic enumeration and clinical validation of all potential combinations of responses over psychometric scales was a preparation of a system of comprehensive therapeutic recommendations. The assessment system used three standardized psychometric scales with 6 to 9 response options each creating a large number of possible profiles of user responses. In all possible combinations of scale responses, therapeutic recommendations and intervention strategies were developed in line with the existing psychological principles and evidence-based practices[1][9].

Since it was understood that the wrong or unsound clinically recommendations may in fact be detrimental to the weakness users, all the generated recommendations were carefully vetted by a licensed psychiatrist who has more than 20 years of clinical experience. The expert critically analyzed every recommendation, replacing inappropriate recommendations, modifying the therapeutic interventions, and adapting them to the evidence-based psychological interventions that fit the Bangladesh setting. This validation process also involved clinical judgement on severity assessment, the correct course of action that should be used in case of crisis situations, and cultural sensitive intervention strategies that do not interfere with local social norms, and family structure.

The emerging system acts like a clinically-based decision support system which will make sure that all the recommendations made available to the users are therapeutically sound and suitable to their particular profile of responses. A mapping that has been verified by experts fuses the capabilities of artificial intelligence with real clinical expertise and forms a secure and dependable base on which AI-assisted mental health support could be built. The system allows assuming appropriate diagnostic screening and maintaining the highest care standards, which oppose the actual lack of mental health specialists in Bangladesh.

7.1.3 Synonym-Enriched Psychometric Scale Dataset

A new approach was designed to connect formal psychometric assessment language and natural expression of the user. Conventional psychometric scales utilize strict response choices which are in a Likert format and these might not reflect the way people express their mental health experiences especially in non-Western societies. To overcome this shortcoming, 50 participants were sampled with the help of data gathered through the use of the established psychometric scales, though, rather than choosing an option among the given ones, participants were proposed to provide answers in the form of free text. This method produced a deep body of natural language variation representing the same basic level of sentiments.

All free-text answers were then checked over by a licensed psychiatrist who cross-tabulated every natural expression with its corresponding formal scale alternative, crossover to form a cross-referential synonym dictionary. This is seen in, e.g., where a scale may OK have "Strongly Agree," the enriched data now has colloquial synonyms of Strongly Agree, like "absolutely," "definitely true," or "I completely feel this way." Such synonym mapping allows the chatbot to understand various linguistic expressions to be accurate, thus being more accessible to any user with various

educational levels, age, and language skills. The mapping, which has been validated by the psychiatrist; clinical accuracy is maintained without any loss of the naturalness of the everyday communication, which is a tremendous step towards the culturally adaptive mental health assessment instruments in Bangladesh and in the like situations.

7.1.4 Systematic Comparative Evaluation of Large Language Models

The article is the first systematic, rigorous, comparison of five state-of-the-art large language models applied to mental health chatbots in resource-constrained environments. The architectural paradigms that the evaluated models (Zephyr 7B Beta[45] , Gemma 2 9B Instruction-Tuned [49], evaluated with two training configurations, Llama 2 7B Instruct [48] and Phi-3 Mini 4k Instruct [46]) represent are dialogue-optimized models, safety-focused instruction-tuned models, general-purpose instruction followers, and compact efficiency-oriented architectures.

All models were evaluated based on a three-level evaluation system containing critical metrics (Semantic Similarity [22], BERTScore F1 [24], Perplexity[13]), important metrics (ROUGE-L[7], Response Length Consistency), and other metrics to fully document all metrics. This standard assessing procedure allowed the comparison of all models fairly because it ensured that the training procedures, hyperparameter choices, and hardware limitations were the same[31]. All of the experiments were run on consumer-grade Tesla P100 GPUs with 4-bit quantization using QLoRA[40], which proved that it could be deployed to low- and middle-income countries.

The comparison analysis provided acute information when choosing models in specific fields. Instruction-tuned models[48][46] performed far better than dialogue-optimized architectures[45], and Zephyr 7B Beta did not meet any of the critical performance benchmarks whereas Llama 2 7B Instruct did. Even the efficiency of their parameters, such as Phi-3 Mini[46], was not enough to be used in safety-critical mental health uses. A study of ablation of the training duration with Gemma 2 9B IT[49] found a counterintuitive trade-off between shorter training and improved semantic alignment and longer training and increased confidence at the expense of semantic drift. These results will serve as practical advice to researchers who will create specific AI systems with limited computational resources [36] [40].

7.1.5 Integration of Clinical Psychometric Scales with Large Language Models

It also created a hybrid model that integrates formal psychometric assessment directly into the conversational AI model, creating a new system of interaction between the structured psychological assessment and the natural language interaction. Three psychometrically assessed tools were used, namely, the Patient Health Questionnaire-4 (PHQ-4) to screen depression and anxiety, the Multidimensional Scale of Perceived Social Support (MSPSS) to measure the network of support, and the Perceived Stress Scale-4 (PSS-4) to assess stress levels. The choice of these

7.2 Limitations and Future Work

Although the system demonstrated promising performance, several limitations must be acknowledged. The 1,000-sample dataset, despite psychiatrist validation, remains relatively small and may not adequately represent the full spectrum of mental health presentations within Bangladesh's population of over 170 million. Additionally, human evaluation was conducted with a limited sample size, which may not fully capture the diversity of user experiences and responses across different demographics and use cases.

A significant constraint is the English-only implementation, which excludes Bangla-speaking users who may not be proficient in English. Future work should prioritize multilingual expansion, incorporating Bangla and other regional languages to ensure broader accessibility and cultural relevance for the diverse linguistic landscape of Bangladesh.

The confidence scores generated by the model exhibited inconsistency across different question types. While psychometric scale questions were humanized to enhance user engagement, the mapping between adapted options and original psychometric scale options was not uniform. Some questions demonstrated strong alignment with established scales, whereas others showed poor correspondence, potentially compromising assessment reliability. Future iterations should focus on developing standardized protocols for adapting psychometric scales while maintaining their psychometric properties, with collaboration from clinical psychologists and psychometricians to validate humanized questions against established instruments.

Furthermore, the model demonstrated hallucination tendencies by generating empathetic responses without sufficient contextual information. This behavior raises concerns about output appropriateness and accuracy, particularly in sensitive mental health contexts where unfounded empathetic statements could mislead users or provide false reassurance. Addressing this requires developing enhanced context-awareness mechanisms and implementing stricter guardrails in future versions.

Validation was limited to initial response quality measures rather than longitudinal tracking of user outcomes, which would better establish clinical efficacy. Future research should implement longitudinal studies tracking user outcomes over extended periods to confirm the system's long-term effectiveness in real-world settings. Additionally, the evaluation relied on a limited number of psychiatrist perspectives, potentially introducing individual biases that could affect quality assessments. Validation should be extended to incorporate diverse clinical viewpoints to enhance generalization and reduce the risk of individual bias in quality measurement.

Building upon these findings, several avenues for future enhancement can significantly improve the system's reliability and clinical utility. Conducting large-scale human evaluations with diverse participant groups across different demographics, cultural backgrounds, and mental health conditions would provide more robust validation and improve result generalizability. Exploring more sophisticated architectures such as Gemma 2 7B or Llama 2 7B [48] could yield additional performance

improvements.

Incorporating multimodal input processing, including voice tone and response timing, would enable the system to better understand user context and improve emotionalunderstanding. Implementing explainability features would allow users and clinicians to understand the reasoning behind the system's responses, fostering trust and transparency. Adaptive testing mechanisms that adjust question difficulty based on user responses could further improve assessment accuracy.

Critical to future development is the implementation of robust human-in-the-loop systems for crisis escalation [18], ensuring that high-risk situations are appropriately managed with professional intervention. Finally, exploring hybrid human-AI approaches where AI assists rather than replaces mental health professionals could maximize benefits while minimizing risks, ultimately creating a more ethical, effective, and clinically appropriate mental health support system for Bangladesh and beyond.

7.3 Conclusion

In this thesis, a refined large language model that offers culturally responsive psychosocial support to Bangladesh users was developed and tested. Llama 2 7B Instruct[48] was selected as the best option among them due to a systematic comparison of five state-of-the-art models[45][49][48][46] because of its deployment-ready performance with semantic similarity of 0.6680, BERTScore F1 of 0.8942, and perplexity of 1.7937. The model was optimized with parameter-efficient QLoRA methods[40], which showed that mental health chatbots of high quality could be trained on consumer-level hardware, which can be deployed in resource-constrained environments. A new sample set of 1,000 psychosocial support conversations were gathered based on direct interviews, and verified by a licensed psychiatrist based on standard psychometric scales, and including different age groups, sex, and occupations. To achieve the quality of therapeutic results, the evaluation scheme was based on both computational measures[22][24][7][13] and clinical validation based on psychiatric assessment scales. The analysis of interpretability showed that the model has a true grasp of the principles of counseling[1] and continuously applies validation, active listening, and empathy responses without relationships that are harmful. The Chain-of-Thought-based reasoning [39] and post-hoc explanation techniques [14] ensured that all decisions made by the model are clear and within the accepted therapeutic models. This project fills a knowledge gap in Bangladesh regarding mental health treatment through the provision of accessible and culturally relevant assistance and clinical requirements to facilitate safe implementation.

Bibliography

[1] C. R. Rogers, *Client-centered therapy: Its current practice, implications, and theory*. Houghton Mifflin Boston, 1951.

[2] J. Weizenbaum, "Eliza—a computer program for the study of natural language communication between man and machine," *Communications of the ACM*, vol. 9, no. 1, pp. 36–45, 1966.

[3] A. Newell, H. A. Simon, et al., *Human problem solving*. Prentice-hall Englewood Cliffs, NJ, 1972, vol. 104.

[4] S. Cohen, T. Kamarck, and R. Mermelstein, "A global measure of perceived stress," *Journal of Health and Social Behavior*, vol. 24, no. 4, pp. 385–396, 1983. DOI: 10.2307/2136404.

[5] G. D. Zimet, N. W. Dahlem, S. G. Zimet, and G. K. Farley, "The multidimensional scale of perceived social support," *Journal of personality assessment*, vol. 52, no. 1, pp. 30–41, 1988.

[6] C. E. Izard, *The Psychology of Emotions*. New York, NY: Plenum Press, 1991. DOI: 10.1007/978-1-4899-0615-1. [Online]. Available: http://dx.doi.org/10.1007/978-1-4899-0615-1.

[7] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*, Barcelona, Spain: Association for Computational Linguistics, Jul. 2004, pp. 74–81. [Online]. Available: https://aclanthology.org/W04-1013/.

[8] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004, pp. 74–81.

[9] M. Cooper, "Essential research findings in counselling and psychotherapy: The facts are friendly," 2008.

[10] K. Kroenke, R. L. Spitzer, J. B. Williams, and B. Löwe, "An ultra-brief screening scale for anxiety and depression: The phq–4," *Psychosomatics*, vol. 50, no. 6, pp. 613–621, 2009.

[11] W. R. Miller and S. Rollnick, *Motivational interviewing: Helping people change*. Guilford press, 2012.

[12] M. D. Hossain, H. U. Ahmed, W. A. Chowdhury, L. W. Niessen, and D. S. Alam, "Mental disorders in bangladesh: A systematic review," *BMC psychiatry*, vol. 14, no. 1, p. 216, 2014.

[13] S. Merity, C. Xiong, J. Bradbury, and R. Socher, "Pointer sentinel mixture models," *arXiv preprint arXiv:1609.07843*, 2016.

[14] F. Doshi-Velez and B. Kim, *Towards a rigorous science of interpretable machine learning*, 2017. arXiv: 1702.08608 [stat.ML]. [Online]. Available: https://arxiv.org/abs/1702.08608.

[15] K. K. Fitzpatrick, A. Darcy, and M. Vierhile, "Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): A randomized controlled trial," *JMIR mental health*, vol. 4, no. 2, e7785, 2017.

[16] A. Vaswani et al., "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[17] B. Inkster, S. Sarda, V. Subramanian, et al., "An empathy-driven, conversational artificial intelligence agent (wysa) for digital mental well-being: Real-world data evaluation mixed-methods study," *JMIR mHealth and uHealth*, vol. 6, no. 11, e12106, 2018.

[18] Z. C. Lipton, "The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery.," *Queue*, vol. 16, no. 3, pp. 31–57, 2018.

[19] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 2019, pp. 4171–4186.

[20] National Institute of Mental Health, Bangladesh, *Mental health survey report*, https://nimh.gov.bd/wp-content/uploads/2021/11/Mental-Health-Survey-Report.pdf, Accessed: 2025-12-03, 2019.

[21] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al., "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[22] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," *arXiv preprint arXiv:1908.10084*, 2019.

[23] S. Wiegreffe and Y. Pinter, "Attention is not not explanation," *arXiv preprint arXiv:1908.04626*, 2019.

[24] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," *arXiv preprint arXiv:1904.09675*, 2019.

[25] V. Zhelezniak, A. Savkov, A. Shen, and N. Hammerla, *Correlation coefficients and semantic textual similarity*, 2019.

[26] A. A. Abd-Alrazaq, A. Rababeh, M. Alajlani, B. M. Bewick, and M. Househ, "Effectiveness and safety of using chatbots to improve mental health: Systematic review and meta-analysis," *Journal of medical Internet research*, vol. 22, no. 7, e16021, 2020.

[27] E. Hekler et al., "Tuning: An approach for supporting healthful adaptation," *Interactions*, vol. 27, no. 2, 2020.

[28] M. A. Islam, S. D. Barna, H. Raihan, M. N. A. Khan, and M. T. Hossain, "Depression and anxiety among university students during the covid-19 pandemic in bangladesh: A web-based cross-sectional survey," *PloS one*, vol. 15, no. 8, e0238162, 2020.

[29] A. Jacovi and Y. Goldberg, "Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness?" *arXiv preprint arXiv:2004.03685*, 2020.

[30] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, and M. Zhou, *Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers*, 2020. arXiv: 2002.10957 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2002.10957.

[31] T. Wolf et al., "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Q. Liu and D. Schlangen, Eds., Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45. DOI: 10.18653/v1/2020.emnlp-demos.6. [Online]. Available: https://aclanthology.org/2020.emnlp-demos.6/.

[32] J. E. Bautista et al., "The completed sdss-iv extended baryon oscillation spectroscopic survey: Measurement of the bao and growth rate of structure of the luminous red galaxy sample from the anisotropic correlation function between redshifts 0.6 and 1," *Monthly Notices of the Royal Astronomical Society*, vol. 500, no. 1, pp. 736–762, 2021.

[33] B. M. Melnyk et al., "Psychometric properties of the short versions of the ebp beliefs scale, the ebp implementation scale, and the ebp organizational culture and readiness scale," *Worldviews on Evidence-Based Nursing*, vol. 18, no. 4, pp. 243–250, 2021.

[34] K. Tsamakis et al., "Covid-19 and its consequences on mental health," *Experimental and therapeutic medicine*, vol. 21, no. 3, p. 244, 2021.

[35] World Health Organization, *Mental Health ATLAS 2020*. Geneva: World Health Organization, 2021, CC BY-NC-SA 3.0 IGO license, ISBN: 9789240036703. [Online]. Available: https://iris.who.int/handle/10665/345946.

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

[37] L. Ouyang et al., "Training language models to follow instructions with human feedback," *Advances in neural information processing systems*, vol. 35, pp. 27 730–27 744, 2022.

[38] X. Wang et al., "Self-consistency improves chain of thought reasoning in language models," *arXiv preprint arXiv:2203.11171*, 2022.

[39] J. Wei et al., "Chain-of-thought prompting elicits reasoning in large language models," *Advances in neural information processing systems*, vol. 35, pp. 24 824–24 837, 2022.

[40] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "Qlora: Efficient finetuning of quantized llms," *Advances in neural information processing systems*, vol. 36, pp. 10 088–10 115, 2023.

[41] M. O. Faruk et al., "Mental illness stigma in bangladesh: Findings from a cross-sectional survey," *Cambridge Prisms: Global Mental Health*, vol. 10, e59, 2023.

[42] M. R. Haque and S. Rubya, "An overview of chatbot-based mobile mental health apps: Insights from app description and user reviews," *JMIR mHealth and uHealth*, vol. 11, no. 1, e44838, 2023.

[43] R. Huque, A. K. Azad, K. Islam, H. U. Ahmed, and M. R. Amin, "Care seeking behavior and treatment gap for mental health conditions in bangladesh: Evidence from the national mental health survey 2019," 2023.

[44] F. Stoehr et al., "Natural language processing for automatic evaluation of free-text answers—a feasibility study based on the european diploma in radiology examination," *Insights into Imaging*, vol. 14, no. 1, p. 150, 2023.

[45] L. Tunstall et al., "Zephyr: Direct distillation of lm alignment," *arXiv preprint arXiv:2310.16944*, 2023.

[46] M. Abdin et al., *Phi-3 technical report: A highly capable language model locally on your phone*, arXiv:2404.14219 [cs.CL], 2024. DOI: 10.48550/arXiv.2404.14219. [Online]. Available: https://arxiv.org/abs/2404.14219.

[47] F. H. Baishakhi, "Level and experience of occupational role among the people with mental illness after receiving occupational therapy services from the selected mental health day centre in bangladesh," Ph.D. dissertation, Bangladesh Health Professions Institute, Faculty of Medicine, the University ..., 2024.

[48] A. Meta, "Llama 3 model card," *GitHub https://github. com/meta-llama/llama-models/blob/main/models/llama3_1/MODEL_CARD. md. Accessed*, vol. 21, 2024.

[49] G. Team et al., "Gemma 2: Improving open language models at a practical size," *arXiv preprint arXiv:2408.00118*, 2024.

[50] DataVisor, *Natural language processing*, https://www.datavisor.com/wiki/natural-language-processing, [Accessed 29-11-2025], 2025.

[51] M. Dechant, E. Lash, S. Shokr, and C. O'Driscoll, "Future me, a prospection-based chatbot to promote mental well-being in youth: Two exploratory user experience studies," *JMIR Formative Research*, vol. 9, no. 1, e74411, 2025.

[52] R. Huque, A. K. Azad, K. Islam, H. U. Ahmed, and M. R. Amin, "Mental health care seeking behavior in bangladesh: Determinants and treatment gaps," *BMC psychiatry*, vol. 25, no. 1, p. 779, 2025.

[53] M. A. Rahaman, "Mental health of handloom weavers in bangladesh: A call for culturally adapted interventions," *Cambridge Prisms: Global Mental Health*, vol. 12, e123, 2025.

DR. ATIQUL HAQ MAZUMDER, MBBS, MD (Psychiatry), PhD (Psychiatry)

- Consultant Psychiatrist and Advisor, Thikana Psychiatric and Drug Addiction Clinic, 3/6, Block B, Humayun Road, Mohammadpur, Dhaka 1207, Bangladesh.
- Postdoctoral Researcher, Departmental of Psychiatry, Research Unit of Clinical Medicine, Faculty of Medicine, University of Oulu, Peltolantie 17, Building PT1, 90220 Oulu, Finland.
- Assistant Professor (Former), National Institute of Mental Health, Dhaka, Bangladesh.
- Mobile (Bangladesh): +8801713423349, Mobile (Finland): +358456943323
- WhatsApp: +358456943323, Website: https://www.thikana.clinic/atiqul.html
- E-mail: atiq10@gmail.com, atiqul.mazumder@oulu.fi

Tuesday, 1 July 2025 11:00 AM Ref: 2025070111001

Letter of Consent

To whom it may concern,

I am providing this letter as a professional courtesy in my capacity as a practicing psychiatrist with over 20 years of clinical experience. Although I am not issuing this on behalf of any government authority or regulatory body, I am providing my consent as a knowledgeable professional to allow the research team to conduct interviews for their academic thesis work.

I understand the nature and purpose of the research, and I have no objection to their conducting these interviews. My permission is being given voluntarily, and the collected information is to be used strictly for academic and research purposes.

Sincerely,

1st July, 2025, Dhaka

(Date and place of signature)

(Electronic signature)

DR. ATIQUL HAQ MAZUMDER,
MBBS, MD (Psychiatry), PhD (Psychiatry)

Consultant Psychiatrist and Advisor, Thikana Psychiatric and Drug Addiction Clinic, Dhaka, Bangladesh.
Postdoctoral Researcher, Research Unit of Clinical Medicine, University of Oulu, Finland.
Assistant Professor (Former), National Institute of Mental Health, Dhaka, Bangladesh.
Mobile (Bangladesh): +8801713423349, Mobile (Finland): +358456943323
WhatsApp: +358456943323, Website: https://www.thikana.clinic/atiqul.html
E-mail: atiq10@gmail.com, atiqul.mazumder@oulu.fi

Declarations:

- BMDC (Bangladesh Medical and Dental Council) registration number: A-29936.
- Practicing in Bangladesh as a medical doctor since 2000 and as a psychiatrist since 2011.
- Involved with Thikana Psychiatric and Drug Addiction Clinic since 2008.
- Affiliated with Department of Psychiatry, University of Oulu, Finland, since 2022.
- Former Assistant Professor in adult psychiatry in Bangladesh (2013-2014).
- Earned a PhD (Doctor of Medicinal Sciences) in Psychiatry from the University of Oulu, Finland in 2022.
- Earned a MD (Doctor of Medicine) in Psychiatry from former Bangabandhu Sheikh Mujib Medical University (BSMMU) in 2011.
- Earned a MBBS (Bachelor of Medicine and Surgery) from Faridpur Medical College (FMC) in 1999.
- Pioneered Nutritional Psychiatry, Holistic Psychiatry, Reverse Meditation, Anti-fragility training and Boredom Training in Bangladesh in 2022.
- Pioneered Telepsychiatry in Bangladesh in 2016.

PROFESSIONAL VERIFICATION CONSENT FORM

Research Project: Developing an Empathetic Conversational System using Fine-Tuned Language Models and Psychometric Evaluation

---

DECLARATION OF PROFESSIONAL VERIFICATION

I, Dr. Atiqul Haq Mazumder [Print Full Name], a licensed and practicing psychiatrist, hereby provide my professional verification and consent for the following research materials developed for the above-mentioned thesis project:

Credentials

- **License Number:** A29936
- **Specialization:** Psychiatry
- **Years of Practice:** 30
- **Institution/Affiliation:** ① University of Oulu, Finland ② Thikana Psychiatric and drug addiction clinic, Bangladesh.
- **Date of Verification:** 23.11.2025

SCOPE OF VERIFICATION

I confirm that I have thoroughly reviewed, evaluated, and verified the following components of this research:

1. Interview Questions

I have examined the interview questions used in this research and confirm that they are:

- [✓] Appropriate for assessing empathetic responses
- [✓] Ethically sound and non-harmful to participants
- [✓] Professionally valid for the stated research objectives
- [✓] Suitable for psychometric evaluation purposes

2. Training Dataset

I have reviewed the dataset consisting of approximately **1,000 conversational samples** and confirm that:

- [✓] The dataset content is appropriate and ethically suitable
- [✓] The data adequately represents empathetic conversational scenarios
- [✓] The dataset is suitable for training language models for empathetic responses
- [✓] No harmful, discriminatory, or clinically inappropriate content is present

3. Psychometric Scales

I have evaluated the psychometric scales used in this research and verify that:

- [✓] The selected scales are clinically valid and appropriate
- [✓] The scales are suitable for measuring empathetic qualities in conversational systems
- [✓] The application of these scales aligns with professional psychological standards
- [✓] The scoring methodology is sound and appropriate

4. Scale Options and Synonyms

I have reviewed the scale options and their synonyms used for evaluation and confirm that:

- [✓] The terminology is clinically accurate
- [✓] The options/synonyms appropriately capture the intended psychological constructs
- [✓] The language used is clear and appropriate for assessment purposes
- [✓] The variations maintain consistent psychological meaning

---

PROFESSIONAL STATEMENT

Based on my professional expertise in psychiatry and psychological assessment, I certify that the materials listed above have been developed and implemented in accordance with:

- Ethical guidelines for psychological research
- Professional standards for psychometric evaluation
- Best practices in mental health and conversational therapy
- Appropriate safeguards for participant wellbeing

I understand that this verification will be used as part of the thesis documentation and research validation process.

CONSENT AND SIGNATURE

I hereby provide my professional consent for the use of my verification in the thesis documentation, acknowledging that:

1. I have independently reviewed all materials mentioned above
2. My verification is based on my professional judgment and expertise
3. I agree to have my credentials cited in the research (if required)
4. I understand this consent may be presented to academic review committees

---

Psychiatrist's Signature: _____

Printed Name: Dr. Atiqul Haq Mazumder

Date: 23.11.2025

Contact Information (Optional):

- Email: atiq10@gmail.com
- Phone: +8801713423349

RESEARCHER ACKNOWLEDGMENT

I, T2420330 [Researcher Name], acknowledge receiving this professional verification from the above-mentioned psychiatrist for my thesis research.

Researcher's Signature: _____, Sajida

Date: 23.11.25

For Official Use:

- Document Reference Number: _____
- Institution: BRAC University
- Ethics Committee Approval (if applicable): _____