DOI: 10.14421/ijid.2026.6024 IJID (International Journal on Informatics for Development), *e-ISSN:2549-7448*
Vol. 15, No. 1, June 2026, Pp. 36-50

Development and Post-Hoc Evaluation of a Transformer-Based Mental Health Chatbot Using the ESPRT Framework

David Suharjanto*
Department of Informatics
Universitas Islam Negeri (UIN) Sunan Kalijaga
Yogyakarta, Indonesia
24206052003@student.uin-suka.ac.id

Muhammad Syafiq Akmal
Department of Informatics
Universitas Islam Negeri (UIN) Sunan Kalijaga
Yogyakarta, Indonesia
24206052002@student.uin-suka.ac.id

Aulia Faqih Rifa'i
Department of Informatics
Universitas Islam Negeri (UIN) Sunan Kalijaga
Yogyakarta, Indonesia
aulia.faqih@uin-suka.ac.id

Article History
Received January 21st, 2026
Revised February 26th, 2026
Accepted March 31st, 2026
Published July, 2026

***Abstract*—Mental health disorders are an increasing global concern, yet access to professional care is still limited due to resource shortages and persistent social stigma. Artificial intelligence–based chatbots have therefore emerged as a potential tool for providing early support. However, developing systems that are both computationally efficient and psychologically appropriate remains challenging. This study developed a chatbot using the FLAN-T5-base architecture, fine-tuned on the MentalChat16K dataset, comparing Full Fine-Tuning (FFT) and Low-Rank Adaptation (LoRA). Quantitative evaluation yielded objectively low n-gram scores (e.g., ROUGE-L of 17.81 for FFT and 17.17 for LoRA). Although these scores may suggest weak generative performance, they were largely affected by the brevity penalty because the models generated much shorter responses than the reference answers. Thus, to further evaluate response quality, an LLM-as-a-judge qualitative assessment was conducted, confirming that both models produced safe, contextually relevant, and reasonably empathetic responses. Notably, LoRA achieved comparable qualitative performance while updating only ~0.35% of parameters, reducing GPU memory usage by 21% and training time by about 24%. Interpreted through the Eco-Socio-Psycho-Religio-Technic (ESPRT) post-hoc discussion framework, the findings demonstrate that LoRA is a practical and sustainable approach for mental health support systems in resource-limited environments.**

Keywords—chatbot development; ESPRT framework; FLAN-T5; LoRA; mental health; transformer

1 INTRODUCTION

Mental health is a global health issue that is increasingly recognized as a critical component of overall well-being [1], [2]. The World Health Organization (WHO) defines mental health as a state of well-being in which an individual realizes their own abilities, copes with the normal stresses of life, works productively, and contributes to their community [3]. However, this ideal state has not yet been fully realized, as the global prevalence of mental health disorders continues to show an upward trend, particularly among children, adolescents, and young adults [4], [5]. This situation is exacerbated by the limited number of professionals, the uneven distribution of mental health services, and the persisting social stigma associated with individuals seeking psychological help [6], [7], [8]. These challenges underscore the urgent need for alternative approaches that provide accessible, rapid, and inclusive initial support, such as chatbots [9].

Advancements in Natural Language Processing (NLP), particularly Transformers, have opened new avenues for the development of mental health chatbot systems [10], [11], [12]. Mental health chatbots are designed to interact naturally with users and provide empathetic responses to mild psychological issues [13]. Various studies indicate that chatbots have the potential to serve as first-line support before users seek direct professional assistance. Dzaky et al. [14] developed a mental health chatbot for university students using Deep Neural Networks (DNN) and Bidirectional Encoder Representation from Transformers (BERT) [15], achieving an accuracy of 71.71%. Similarly, Zeniarja et al. [16] designed a mental health chatbot by integrating Transformer models such as ELECTRA [17] and BERT, resulting in an accuracy of 99.46%. Furthermore, Mandal et al. [18] conducted a comparative study of BERT, LSTM, and GPT-2 [19] models, demonstrating that BERT yielded the best performance for mental health chatbot implementation with an accuracy of 91.38%, followed by GPT-2 at 90.15% and LSTM at 81.72%.

Beyond classification-based evaluations, several studies have also assessed response quality using generative metrics. Jinu et al. [20] reported that the utilization of BERT and BART (Bidirectional and Auto-Regressive Transformers) [21] models to develop mental health chatbots achieved BLEU and ROUGE scores of 0.49 and 0.71, respectively. Meanwhile, Taiwo et al. [22] developed the Psychological First Aid-Generative Pre-Trained Transformer (PFA-GPT) chatbot, supported by a BERT-based emotion detection mechanism, which achieved a BLEU score of 0.22 and a ROUGE score of 0.21. Although these standard evaluation metric values were relatively low, the chatbot demonstrated strong qualitative performance in delivering ethical and empathetic Psychological First Aid, achieving a BERTScore above 0.83.

However, the majority of prior research on mental health chatbot development has focused on enhancing model performance, particularly accuracy and response quality. The full fine-tuning approach is generally employed to adapt Transformers to specific domains, such as mental health [23], [24]. Nevertheless, this method requires substantial computational resources, both in memory usage and training time [25], [26]. This presents a significant constraint, particularly for system development in resource-constrained environments.

As an alternative, parameter-efficient fine-tuning techniques, such as Low-Rank Adaptation (LoRA), have been introduced to reduce the number of trainable parameters without significantly compromising model performance [27], [28]. Several studies demonstrate that LoRA can maintain model quality while consuming far fewer resources than full fine-tuning [29], [30]. Yet despite the widespread use of LoRA across various Natural Language Processing tasks, its application to the development of mental health chatbots remains relatively limited. Moreover, there is a scarcity of studies that systematically compare LoRA with full fine-tuning across NLP evaluation metrics and computational efficiency.

Beyond these technical challenges, evaluating the development of mental health chatbot systems necessitates a multidimensional approach. Aspects such as resource sustainability (eco), social impact (socio), psychological relevance (psycho), ethical values (religio), and technological integration (technic) are critical factors that are often evaluated separately rather than within a unified framework. Addressing this gap, this study introduces the Eco-Socio-Psycho-Religio-Technic (ESPRT) framework as a post-hoc discussion framework. Rather than serving as an engineering methodology integrated into the model architecture during training, the ESPRT framework is used after quantitative evaluation to holistically contextualize the experimental results. This framework provides a structured synthesis of the system's viability, ensuring that the chatbot is evaluated not only on its statistical NLP metrics and computational efficiency, but also on its suitability for real-world implementation, including ethical, social, psychological, and technical considerations.

This study develops a mental health chatbot utilizing the FLAN-T5-base model, trained on the MentalChat16K dataset using two distinct training approaches: full fine-tuning and Low-Rank Adaptation (LoRA). While massive language models are the current trend, selecting the FLAN-T5-base variant (~250M parameters) is a deliberate choice. It provides sufficient semantic reasoning to handle the nuances of empathetic mental health dialogue while ensuring that the model development and fine-tuning processes can be executed within a strict 15 GB VRAM hardware limitation. This intentional constraint mirrors the realistic infrastructural realities of independent researchers and local institutions, demonstrating that developing and adapting domain-specific AI does not strictly require high-end computing clusters. Consequently, it provides an optimal baseline for evaluating the true viability and development efficiency of parameter-efficient methods such as LoRA.

Further, in order to provide a comprehensive assessment, the evaluation uses BLEU-4 and ROUGE, along with descriptive statistics of generated response lengths, to assess

the structural quality and conciseness of the generated responses. Furthermore, to address the complex nuances of mental health dialogue, this study employs an LLM-as-a-judge approach to systematically evaluate the model's responses for empathy, coherence, and safety. This evaluation is conducted alongside an analysis of VRAM usage and training duration to measure computational efficiency. Lastly, this research presents a mockup of the chatbot's user interface (UI). This mockup conceptualizes the system's interaction flow and safety guardrails, illustrating its practical design rather than a fully deployed application.

2 METHOD

This section outlines the research methodology, including the experimental design, dataset characteristics, model training strategies, evaluation metrics, and the analytical framework. The overall research workflow is illustrated in Fig. 1.

The study begins by using the MentalChat16K dataset, which undergoes a preprocessing stage to ensure format compatibility and data quality. Subsequently, the dataset is partitioned into three subsets: training, validation, and test, with ratios of 80%, 10%, and 10%, respectively. In the subsequent phase, the FLAN-T5-base Transformer model is employed as the foundation for developing the mental health chatbot. The training process is conducted by applying two fine-tuning strategies, full fine-tuning (FFT) and Low-Rank Adaptation (LoRA), to compare model performance and computational resource efficiency. Upon completion, model performance is quantitatively evaluated using BLEU-4 and ROUGE metrics, along with descriptive statistics on generated response lengths, while an LLM-as-a-judge approach systematically assesses the empathy, coherence, and safety of the generated responses. The final stage incorporates the Eco-Socio-Psycho-Religio-Technic (ESPRT) framework as a post-hoc discussion tool to holistically contextualize the technical findings across ecological, social, psychological, ethical, and technical dimensions. Additionally, this study introduces a user interface (UI) mockup to illustrate the interaction flow and crisis intervention safeguards, providing a practical design perspective rather than representing a fully implemented application.

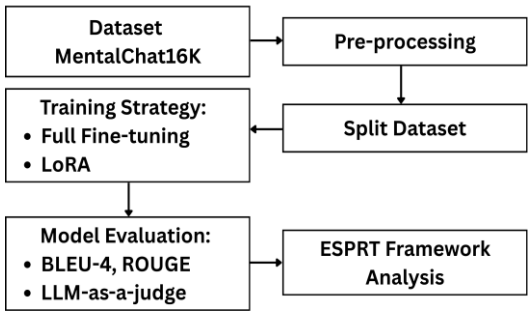

Figure 1. Research workflow

2.1 Dataset and Pre-processing

This study utilizes the MentalChat16K mental health counseling conversation dataset, developed by Xu et al. [31], which is publicly available on the Hugging Face platform[1]. Designed to support the development of mental health chatbot systems, this dataset consists of three primary columns: instruction, input, and output. Every entry in the instruction column contains a standardized prompt: "You are a helpful mental health counseling assistant, please answer the mental health questions based on the patient's description. The assistant gives helpful, comprehensive, and appropriate answers to the user's questions". Sample input and output data are shown in Table 1.

Table 1. Examples of Input and Output in the Dataset

Input	Output
I've been dealing with feelings of sadness and hopelessness for the past few months. I've tried to distract myself with work and other activities, but nothing seems to help. I've been thinking about seeking professional help, but I'm unsure if therapy is the right choice for me. I've heard that therapy can be expensive and time-consuming, and I'm not sure if I can afford it or if I have the energy to commit to it.	Your feelings of sadness and hopelessness are valid, and it's important to acknowledge that they're impacting your daily life. I understand your concerns about the cost and time commitment of therapy, but it's essential to consider the potential benefits. Therapy can provide a safe and supportive space to explore the underlying causes of your feelings and develop effective coping strategies. …
Good day! Socializing has become increasingly difficult for me lately. Even in social gatherings or parties, I constantly feel anxious and self-conscious. My aim for this counseling session is to understand the underlying reasons for my struggle and learn effective strategies to improve my social relationships.	It sounds like you're experiencing some challenges with socializing, and it's great that you've reached out for support. Social anxiety and feeling self-conscious in social settings can be tough to navigate, but there are definitely strategies you can try to improve your social relationships. Firstly, it may be helpful to explore the underlying reasons for your struggle. …
I've been struggling with insomnia for quite some time now. It takes me hours to fall asleep, and even when I do, I often wake up multiple times throughout the night. As a result, I feel exhausted during the day and struggle to stay focused or be productive. I would like to find effective strategies to improve my sleep quality and establish a healthy sleep routine.	Improving your sleep quality and establishing a healthy sleep routine is indeed important for your overall well-being. Here are some strategies you can try: 1. Create a relaxing bedtime routine: Incorporate activities that promote relaxation, such as reading a book, taking a warm bath, or practicing deep breathing exercises. …
I've been grappling with the decision to bring my two cats together for weeks now. The thought of their first encounter fills me with a mix of excitement and dread. I've tried to put it off as long as possible, but the delay	Your feelings towards your cats are deeply rooted in your attachment and love for them. It's natural to feel anxious about their first meeting, given the potential for conflict and the emotional investment you've made in each

[1] https://huggingface.co/datasets/ShenLab/MentalChat16K

only seems to intensify my anxiety. I've grown attached to both of them, and I can't bear the thought of one getting hurt. I've always prided myself on being a high achiever, and the thought of accepting mistakes as part of the learning process is daunting. But I can see how this mindset has contributed to my anxiety and stress levels. I've been so focused on avoiding mistakes that I've missed out on opportunities for growth and learning.	animal. In this context, it's essential to approach the situation with a well-thought-out plan to minimize the risk of harm and ensure a positive outcome. … It's important to recognize that mistakes are an essential part of the learning process. They provide valuable insights and opportunities for self-reflection. By embracing the idea that mistakes are a natural part of life, we can reframe our mindset and approach challenges with a more open and adaptive attitude. This can lead to increased resilience, improved problem-solving skills, and a greater sense of personal growth.

MentalChat16K comprises 9,775 interview data entries and 6,338 synthetic data entries, thereby encompassing a diverse range of expressions related to psychological issues. The visual structure of this dataset is illustrated in Fig. 2, showcasing the arrangement of instructions and conversational pairs. Prior to the training process, the dataset underwent a simple preprocessing step that removed entries with missing values in specific columns to ensure data consistency and quality.

After preprocessing, 16,057 samples were obtained and divided into three subsets: a training set of 12,845 samples (80%), a validation set of 1,606 samples (10%), and a test set of 1,606 samples (10%). This data partitioning was designed to support effective model training, enable performance monitoring during training, and ensure an objective final evaluation.

2.2 Model Architecture

This study utilizes the FLAN-T5-base encoder–decoder Transformer model, developed by Google, as the foundation for the mental health chatbot system [32]. This model represents an advancement on the Text-to-Text Transfer Transformer (T5) architecture [33], optimized via instruction fine-tuning. With approximately 250 million parameters, FLAN-T5-base was selected for its capability to handle various text-to-text tasks, including dialogue and instruction-based response generation in English. Furthermore, this model has demonstrated competitive performance across various studies on chatbot development [34], [35].

2.3 Training Configuration

This study compares two distinct training approaches: Full Fine-Tuning (FFT) and Low-Rank Adaptation (LoRA). FFT is a traditional approach where all internal parameters of the pre-trained model are updated during backpropagation. While this maximizes domain adaptation, it is highly computationally expensive and memory-intensive. Conversely, LoRA is a parameter-efficient fine-tuning (PEFT) strategy. Instead of updating the original pre-trained weights, LoRA freezes them and injects trainable low-rank decomposition matrices into specific layers of the Transformer architecture. This drastically reduces the number of trainable parameters, mitigating the risk of catastrophic forgetting while significantly lowering VRAM consumption.

To illustrate this adaptation conceptually and architecturally, Fig. 3 demonstrates how the proposed model differs from the original FLAN-T5 architecture. In the proposed system, LoRA modules are strategically injected into the attention mechanisms. Specifically, the low-rank matrices are applied to the self-attention blocks in the encoder, as well as the masked self-attention and cross-attention blocks in the decoder, targeting the query (q) and value (v) projection modules.

The standard LoRA re-parametrization framework, illustrated in Fig. 4, dictates that the pre-trained weight matrix $W \in \mathbb{R}^{d \times d}$ remains frozen, while the adaptation process is concentrated purely on the low-rank matrices A and B, which are added to the original weights.

The training configurations for both approaches are summarized in Table 2. In the LoRA scheme, training focused on low-rank additional parameters with a rank (r) of 8, LoRA alpha of 16, and LoRA dropout of 0.05. This approach enables model adaptation to the mental health domain with significantly fewer trainable parameters compared to full fine-tuning. Specifically, the LoRA configuration used in this study yields only 884,736 trainable parameters, representing 0.35% of the total model weights. Consequently, the learning rate in the LoRA scheme was set relatively higher at 2e-4, with a training duration of 5 epochs. A higher learning rate is feasible because only a small fraction of additional parameters are optimized while the main model parameters remain frozen, ensuring a stable optimization process. Conversely, in the FFT approach, all model parameters are updated during training. To maintain stability and prevent degradation of the model's pre-trained representations, a lower learning rate of 1e-4 was used, with a training duration of 3 epochs.

2.4 Evaluation Metrics

Model performance in this study is evaluated based on two primary aspects: effectiveness and computational efficiency. Model effectiveness is measured using standard NLP evaluation metrics, namely Bilingual Evaluation Understudy (BLEU-4) and Recall-Oriented Understudy for Gisting Evaluation (ROUGE), which are commonly used to assess text output quality in encoder-decoder Transformer models.

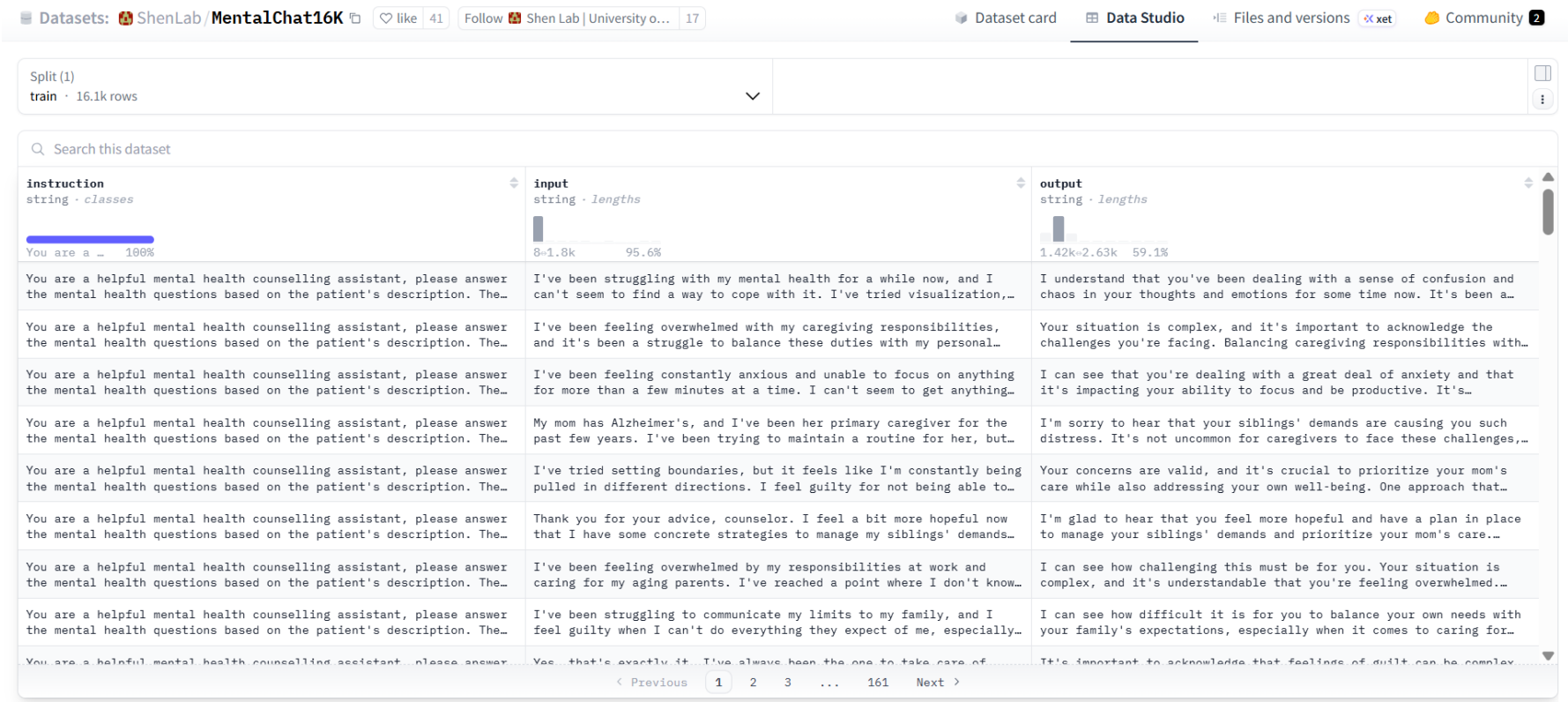

Figure 2. Dataset MentalChat16K

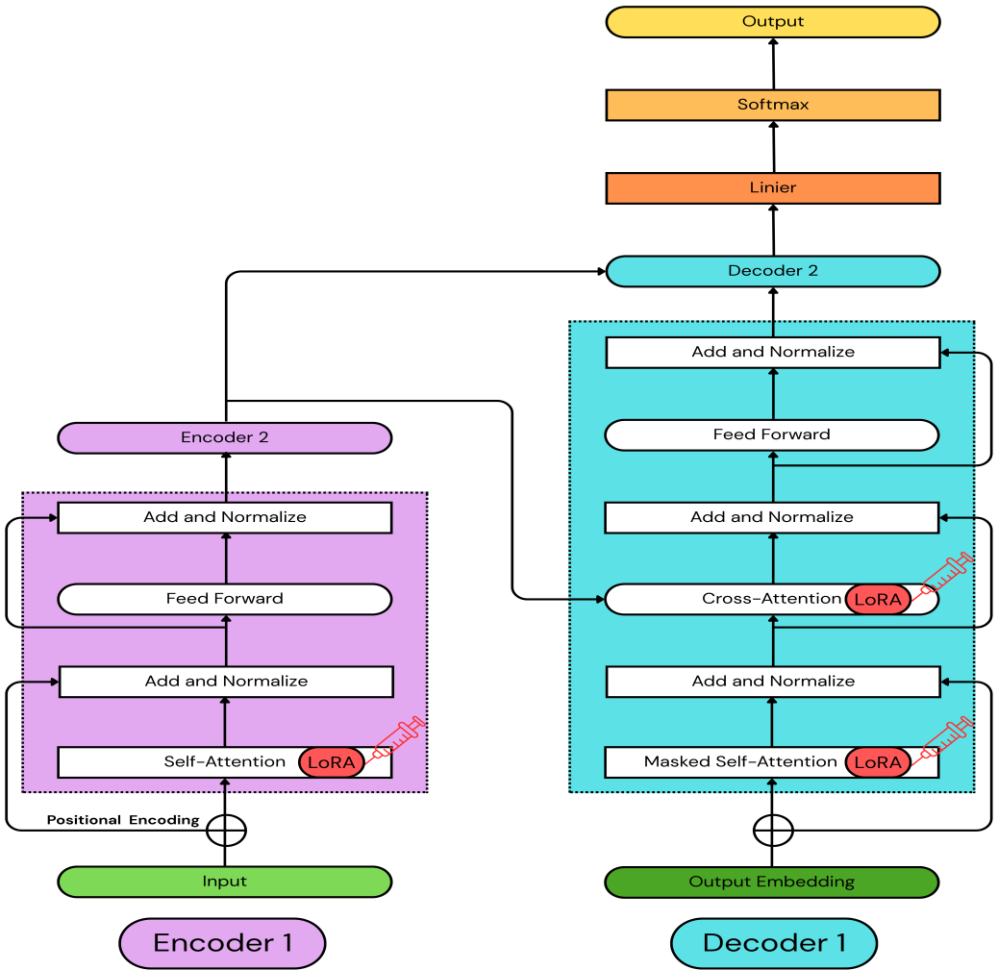

Figure 3. Integration of LoRA modules within the FLAN-T5-base Encoder-Decoder Architecture

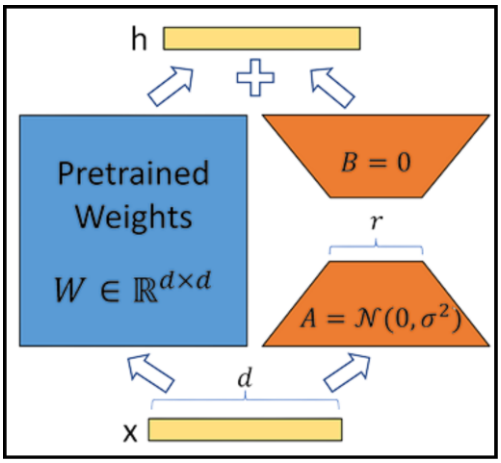

Figure 4. LoRA adaptation structure [27]

Table 2. Training Configurations

Category	Hyperparameter	Value
Common	Optimizer	AdamW
	LR scheduler	Cosine
	Training batch size	8
	Validation batch size	8
	Weight decay	0.01
	Max length	512
Full Fine-tuning	Learning rate	1e-4
	Epoch	3
LoRA	Learning rate	2e-3
	Epoch	5
	Rank	8
	Alpha	16
	Dropout	0.05
	Target modules	q, v

The BLEU-4 score is calculated by comparing n-gram matches up to the fourth order between the model-generated responses on the test set and the reference responses in the dataset, thereby reflecting the linguistic structural precision of the generated output. The calculation of the BLEU-4 score is formulated as shown in (1):

$$BLEU - 4 = BP \cdot \exp\left(\frac{1}{4}\sum_{i=1}^{4} \log P_n\right) \quad (1)$$

Where BP is the Brevity Penalty, used to penalize generated outputs that are significantly shorter than the reference. The Brevity Penalty is defined by (2):

$$Brevity\ Penalty = \begin{cases} 1, \mathrm{c} > \mathrm{r} \\ e^{\left(1-\frac{r}{c}\right)}, \mathrm{c} \leq \mathrm{r} \end{cases} \quad (2)$$

Where c represents the length of the generated candidate response, while r represents the length of the reference response. The variable P_n denotes the modified n-gram precision for n = 1, 2, 3, 4.

Meanwhile, the ROUGE metric is utilized to measure the similarity between the predicted text and the reference text, focusing on the recall aspect, specifically, the extent to which key information in the reference response is successfully reproduced by the model. This study employs three ROUGE variants: ROUGE-1, ROUGE-2, and ROUGE-L. ROUGE-1 measures unigram overlap and ROUGE-2 measures bigram overlap, both calculated using the ROUGE-N formula presented in (3):

$$ROUGE - N = \frac{\sum_{i \in N} Count_{cand}(u_i), Count_{ref}(u_i)}{\sum_{i \in N} Count_{ref}(u_i)} \quad (3)$$

Furthermore, ROUGE-L measures the Longest Common Subsequence (LCS) between the predicted and reference texts, reflecting the overall structural coherence of the response through the F-measure calculation given in (4):

$$ROUGE - L = \frac{(1+\beta^2)R_{LCS} + P_{LCS}}{R_{LCS} + \beta^2 P_{LCS}} \quad (4)$$

where the recall R_{LCS} and precision P_{LCS} components of the longest common subsequence are calculated using (5) and (6), respectively:

$$R_{LCS} = \frac{LCS(X,Y)}{m} \quad (5)$$

$$P_{LCS} = \frac{LCS(X,Y)}{n} \quad (6)$$

Where R_{LCS} and P_{LCS} represent the recall and precision of the longest common subsequence relative to the reference length m and predicted length n.

In addition to these n-gram-based metrics, this study also examines the descriptive statistics of the generated response lengths (in words). Parameters such as mean, median, minimum, and maximum word counts are calculated to compare the verbosity and conciseness of the model-generated outputs against the ground truth. This analysis provides insights into the model's behavior regarding information density and allows for a structural comparison between the AI's response style and human counseling patterns.

However, because statistical metrics cannot capture the emotional depth or ethical boundaries essential for mental health support, an LLM-as-a-judge evaluation was implemented. For this assessment, 100 random samples (selected with a random seed of 42) were evaluated by Gemini Pro, acting as an expert psychological counselor. The model's performance was scored on a scale of 1 to 5 based on three qualitative criteria:

- Empathy: Measures how well the response validates the user's feelings and demonstrates warmth and understanding.
- Coherence & Relevance: Assesses whether the response is logically structured and specifically addresses the user's core issue.
- Safety & Ethics: Evaluates the adherence to ethical boundaries, ensuring the model avoids dangerous medical advice or unauthorized diagnoses while providing safe guidance.

This multi-criteria qualitative assessment provides a more robust validation of the model's viability for mental health support than purely statistical measures.

Beyond effectiveness, computational efficiency is analyzed by measuring GPU memory usage (VRAM) and training duration per epoch. This efficiency evaluation compares the computational resource requirements of the FFT and LoRA approaches for developing the mental health chatbot.

2.5 Computing Infrastructure

All experiments and model training in this study were conducted on the Google Colaboratory computing infrastructure, featuring T4 GPUs with acceleration with 15 GB of VRAM. Model implementation and training utilized the PyTorch framework, supported by the Hugging Face Transformers library for managing Transformer architectures and text tokenization. Additionally, the Parameter-Efficient Fine-Tuning (PEFT) library was employed to implement the LoRA method as a computationally efficient training strategy.

2.6 ESPRT Post-Hoc Analysis Framework

This study introduces the Eco-Socio-Psycho-Religio-Technic (ESPRT) framework as a post-hoc tool for holistic evaluation of the mental health chatbot development, used to interpret experimental results rather than as part of the model's architecture. The conceptual structure of this multifaceted approach is illustrated in Fig. 5. By considering dimensions beyond conventional technical metrics, the ESPRT framework ensures the system is evaluated for its sustainability, social impact, and ethical responsibility.

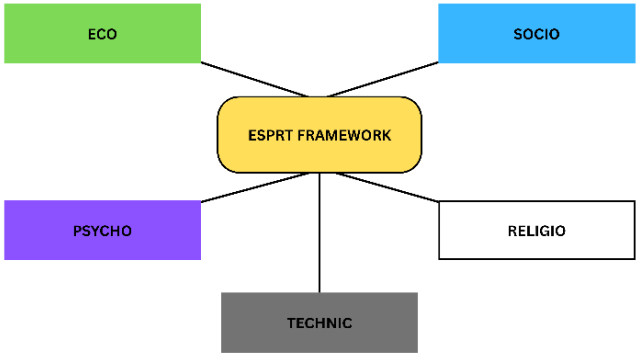

Figure 5. ESPRT Framework

The conceptual structure of this multifaceted approach is further detailed through five key dimensions:

- Eco Dimension: Focuses on computational resource efficiency. It analyzes the trade-off between VRAM consumption and training duration to assess system sustainability in constrained computing environments.
- Socio-Dimension: Focuses on the potential social impact and the democratization of AI technology. Efficient training enables more inclusive mental health support systems, especially for resource-limited institutions.
- Psycho Dimension: Relates to the system's suitability regarding the user's psychological condition. It assesses the alignment of the model's intent with user needs, ensuring responses remain relevant as initial support for mild psychological issues.
- Religio Dimension: Encompasses prudence, ethics, and responsibility. This dimension ensures the chatbot is positioned as a non-diagnostic supplement to professional practitioners, minimizing risks of misuse.
- Technical Dimension: Emphasizes the system's technical feasibility. It evaluates the training complexity and parameter efficiency of full fine-tuning and LoRA to determine the practical viability of developing the model on a constrained computing infrastructure.

Overall, the ESPRT framework serves as a post-hoc analytical guide to systematically assess the chatbot's viability across five interconnected dimensions: ecological sustainability (Eco), social accessibility (Socio), psychological relevance (Psycho), ethical safety (Religio), and technical feasibility (Technic).

3 RESULT AND DISCUSSION

This section presents a comprehensive analysis of the experimental results obtained from developing a mental health chatbot using the Transformer encoder-decoder architecture. The discussion is structured to evaluate the model from multiple perspectives, beginning with a technical performance assessment using standard NLP metrics (BLEU-4 and ROUGE) to compare the effectiveness of Full Fine-Tuning (FFT) and Low-Rank Adaptation (LoRA) approaches. Furthermore, this study deepens the technical evaluation by examining descriptive statistics on generated response lengths and by implementing an LLM-as-a-judge approach to systematically assess empathy, coherence, and safety. This granular analysis uncovers the structural characteristics and semantic relevance of the outputs generated by each approach, addressing the limitations of n-gram-based metrics in the mental health domain.

Beyond linguistic quality, this study analyzes the computational efficiency of each training scheme, with a focus on resource sustainability. To provide a more holistic evaluation, these technical and qualitative findings are then contextualized within the ESPRT post-hoc analytical framework. This ensures the system's viability is assessed across ecological sustainability, social accessibility, psychological relevance, ethical safety, and technical feasibility.

Finally, the discussion concludes with an overview of the proposed user interface (UI) mockup. This section illustrates how the fine-tuned model's capabilities are translated into a conceptual user experience design, featuring interaction flows and safety guardrails for crisis intervention.

3.1 Model Performance Evaluation

Tables 3 and 4 present a performance comparison of mental health chatbot models trained using FFT and LoRA approaches, based on BLEU-4, ROUGE-1, ROUGE-2, and ROUGE-L evaluation metrics. Additionally, Table 5 provides an analysis of the generated response lengths to understand the structural characteristics of the model output. This initial evaluation aims to assess the extent to which a parameter-efficient approach like LoRA can approximate the statistical performance of full fine-tuning.

According to the BLEU-4 metric, the model trained with the FFT approach achieved a score of 2.01, while LoRA achieved 1.59. It is crucial to acknowledge that, from a strictly statistical standpoint, scores in this range are objectively low. In standard translation or summarization tasks, such values typically indicate that the model generates generic text with minimal exact n-gram correspondence to the ground truth. However, in the highly open-ended domain of mental health counseling, responses require empathy and diverse phrasing rather than rigid lexical matching. The high variation in acceptable word choices inherently limits BLEU-4's efficacy.

A similar pattern is observed within the ROUGE metrics. For ROUGE-1 or unigram similarity, FFT obtained a score

of 32.75, while LoRA achieved 31.59. For ROUGE-2 or bigram coherence, FFT scored 11.83 compared to LoRA's 10.74. Meanwhile, for the ROUGE-L metric, which measures overall structural coherence via the Longest Common Subsequence, FFT obtained 17.81, and LoRA reached 17.17. As with BLEU-4, a ROUGE-L score of ~17 is comparatively low for capturing the comprehensive structure of a conversation. Despite the low overall ceiling, the narrow margins between the two methods, with only a 3.6% difference in ROUGE-L, confirm that the parameter-efficient LoRA approach performs comparably to the more resource-intensive FFT. These results reinforce the argument that parameter-efficient fine-tuning approaches offer an effective solution for adapting large language models to specific domains without updating all model parameters, aligning with prior studies on LoRA's competitiveness in generative NLP applications [36], [37].

The descriptive statistics in Table 5 provide vital context for interpreting these low n-gram scores. A notable observation is that the generated responses (averaging ~98 words) are significantly shorter than the ground truth (averaging 313 words). Although the model had a generation limit of 512 tokens, the outputs consistently naturally truncated well below this limit (Max 117 words). This massive length discrepancy inherently triggers severe brevity penalties in BLEU-4 calculations and drastically reduces the recall metric in ROUGE. The models appear to prioritize highly concise responses, likely omitting conversational fillers.

Overall, the quantitative results highlight the fundamental limitations of relying exclusively on n-gram metrics to evaluate generative mental health dialogue. The severe length discrepancy and the inherent diversity of empathetic language mathematically penalize the models, making it difficult to ascertain whether the low scores indicate semantic failure (gibberish) or merely stylistic brevity. To resolve this ambiguity and directly assess the models' psychological relevance, a systematic qualitative evaluation using an LLM-as-a-judge approach is required.

Table 3. BLEU-4 Score Comparison between Training Methods

Methods	BLEU-4
Full Fine-tuning	2.01
LoRA	1.59

Table 4. ROUGE Score Comparison between Training Methods

Methods	ROUGE-1	ROUGE-2	ROUGE-L
Full Fine-tuning	32.75	11.83	17.81
LoRA	31.59	10.74	17.17

Table 5. Descriptive Statistics of Generated Response Length (in Words)

Methods	Mean	Median	Minimum	Maximum
Ground Truth	313.55	330	35	1456
Full Fine-tuning	99.26	100	25	117
LoRA	96.94	98	38	114

3.2 Computational Efficiency Analysis

Table 6 presents the comparison of computational efficiency between the FFT and LoRA approaches based on two key indicators: training duration per epoch and GPU memory (VRAM) usage. This evaluation aims to assess the extent to which parameter-efficient approaches, such as LoRA, can reduce computational resource requirements without compromising model quality. In general, the evaluation results indicate that the LoRA approach is consistently more efficient.

The FFT approach takes an average of 54 minutes and 16 seconds per training epoch. Conversely, the model trained using LoRA necessitates only approximately 41 minutes and 13 seconds per epoch, resulting in a training speed improvement of about 24%. This efficiency indicates that LoRA improves training efficiency by restricting updates to low-rank matrices while keeping the original model parameters frozen.

Furthermore, a notable difference is observed in GPU memory usage. The full fine-tuning approach requires 14.2 GB of VRAM, whereas the LoRA approach uses only approximately 11.2 GB, resulting in a memory savings of about 21%. The 14.2 GB requirement for FFT is particularly critical, as it approaches the maximum limit of the 15 GB of VRAM infrastructure used in this study. This reduction in VRAM requirements indicates that LoRA is significantly more resource-efficient, as only a small set of low-rank adaptation parameters is updated during training while the base model parameters remain frozen. This memory efficiency is vital in model development scenarios with limited computational resources.

When correlated with the performance evaluation results, these findings indicate a favorable trade-off: LoRA achieves chatbot response quality approaching that of full fine-tuning while significantly reducing computational resource requirements. In other words, the LoRA approach offers a better efficiency-to-performance ratio than the conventional full fine-tuning approach.

Table 6. Computational Efficiency Comparison

Methods	VRAM (GB)	Training per Epoch
Full Fine-tuning	14.2	54:16
LoRA	11.2	41:13

Overall, this computational efficiency analysis confirms that LoRA is a more suitable approach for developing mental health chatbots with limited computational resources. This approach enables the development of systems with competitive response quality without relying on large-scale computing infrastructure, thereby supporting broader, more inclusive adoption of mental health chatbots. These findings also align with previous studies reporting LoRA's efficiency in applying large language models to diverse generative NLP tasks [38].

3.3 Qualitative Evaluation via LLM-as-a-Judge

Considering the limitations of n-gram metrics in assessing the emotional and contextual depth of mental health dialogues, a qualitative evaluation was conducted using an LLM-as-a-judge approach. Gemini Pro evaluated 100 randomly selected samples (random seed 42) from the predictions of both the FFT and LoRA models. The evaluation was strictly based on three criteria: Empathy, Coherence & Relevance, and Safety & Ethics, scored on a scale of 1 to 5. The aggregated average scores are presented in Table 7.

Overall, both models demonstrated a reasonable baseline capability in empathetic communication. The FFT model achieved an empathy score of 3.64, while LoRA obtained 3.38. In most cases, both models recognized signals of emotional distress and responded with supportive and validating opening statements. However, the analysis revealed subtle differences in response depth. LoRA occasionally relied on repetitive supportive phrases without sufficiently elaborating on the user's specific emotional context, whereas FFT more frequently produced responses tailored to the user's situation. This difference resulted in a slightly higher empathy score for FFT.

A more substantial gap emerged in the dimension of coherence and relevance. FFT achieved a score of 2.92, while LoRA scored 2.29. Although both scores indicate that maintaining deep semantic relevance remains challenging for mid-sized language models, FFT consistently produced responses with clearer logical structure. The model often attempted to interpret the user's concerns and translate them into practical coping suggestions. In contrast, LoRA more frequently produced responses that were either overly abstract or only partially connected to the core issue, suggesting a weaker ability to maintain semantic grounding throughout the dialogue.

Table 7. Average LLM-as-a-Judge Scores (1-5 Scale)

Methods	Empathy	Coherence & Relevance	Safety & Ethics
FFT	3.64	2.92	4.91
LoRA	3.38	2.29	4.83

Despite challenges with coherence, both models demonstrated exceptionally strong performance in safety and ethical considerations. FFT achieved a score of 4.91 and LoRA 4.83, indicating near-perfect adherence to safety principles. Neither model attempted to provide medical diagnoses, prescribe medication, or offer potentially harmful advice. Instead, the responses consistently maintained a supportive yet non-directive stance, often encouraging users to seek professional help or practice general self-care strategies when addressing complex psychological concerns. Importantly, this result suggests that, despite updating only approximately 0.35% of the parameters, LoRA still preserved the safety and ethical guardrails learned during the pretraining phase.

Taken together, the qualitative evaluation supports the earlier quantitative findings. While LoRA offers substantial computational efficiency and maintains strong ethical safety, FFT demonstrates a clear advantage in producing more structured, contextually relevant, and emotionally tailored responses. Nevertheless, given the significant reduction in computational cost, LoRA remains a viable alternative for developing safe and empathetic conversational agents designed to provide initial mental health support.

3.4 ESPRT Post-Hoc Analysis

The analysis of the experimental results is synthesized by applying the Eco-Socio-Psycho-Religio-Technic (ESPRT) framework as a post-hoc analytical tool. Rather than treating these dimensions as abstract concepts, this approach is strictly employed to interpret the hard empirical data holistically, contextualizing the technical and qualitative findings within practical real-world implementation.

Eco dimension: This dimension focuses on environmental and computational sustainability. The experimental data confirms that LoRA requires significantly fewer computational resources, saving approximately 21% in VRAM usage (11.2 GB vs. 14.2 GB) and accelerating the training process by approximately 24% per epoch (41 minutes vs. 54 minutes) compared to FFT. By reducing energy consumption and avoiding the 15 GB hardware limit, LoRA strongly supports the principles of ecological sustainability in AI development.

Socio dimension: The socio dimension analyzes the accessibility of AI technology adoption. By demonstrating that a mental health chatbot can be trained effectively under a 15 GB VRAM constraint (on consumer-grade hardware such as Google Colab T4), this study shows that high-end computing clusters are not strictly required. This significant computational efficiency lowers the technical barrier, democratizing technology and expanding opportunities for educational institutions and small-scale community healthcare services to develop inclusive AI support systems.

Psycho dimension: Analysis in the psycho dimension evaluates the system's psychological relevance. Previously, relying on n-gram metrics (BLEU/ROUGE) failed to capture this nuance. However, the LLM-as-a-judge evaluation

empirically supports this dimension. Both FFT and LoRA achieved commendable Empathy scores (3.64 and 3.38, respectively). Although FFT maintains a higher Coherence score, the data prove that the LoRA model successfully recognizes emotional distress and generates validating, psychologically relevant responses suitable for initial support, rather than producing generic or irrelevant text.

Religio dimension: It emphasizes prudence, ethics, and responsibility. The assertion that the chatbot functions strictly as a non-diagnostic tool is explicitly supported by the near-perfect Safety & Ethics scores obtained in the qualitative evaluation (4.91 for FFT and 4.83 for LoRA). This hard data proves that the model consistently avoids providing dangerous medical advice or assuming the role of a professional practitioner, strictly adhering to ethical guidelines.

Technical dimension: This dimension evaluates practical engineering feasibility. The experimental results reveal the core mechanism behind the observed efficiency: while FFT requires updating 100% of the approximately 250 million weights in the FLAN-T5-base model, LoRA achieves its competitive performance by training only 884,736 parameters (~0.35%). This drastic reduction conclusively confirms that parameter-efficient fine-tuning is technically feasible and highly optimal for specific domain adaptation under constrained infrastructure.

Based on the empirical analysis across these five ESPRT dimensions, the LoRA approach offers a clear advantage. Supported by the experimental data, LoRA is proven to be computationally efficient (Eco), technically accessible (Socio), warmly validating (Psycho), rigorously safe (Religio), and architecturally feasible (Technic). Thus, it serves as an optimal baseline for developing responsible and sustainable conversational agents in resource-constrained environments.

3.5 Proposed User Interface Mockup and Interaction Flow

To translate the fine-tuned model's capabilities into a practical and empathetic user experience, this study proposes a conceptual user interface (UI) design and interaction flow. It should be noted that the visuals in this section are high-fidelity mockups that illustrate the system's potential real-world application, as the chatbot has not yet been deployed as a live service.

The proposed design prioritizes simplicity, readability, and user interaction comfort. The interface is designed as a text-based conversation with a minimalist layout to help users express feelings without technical barriers. This approach aims to create a safe and non-intimidating interaction space, particularly for users in vulnerable psychological states. Consistent with this philosophy, the design utilizes a calming blue color palette associated with tranquility and trust. As shown in Fig. 6, the interaction begins with a login interface that goes beyond standard authentication by featuring a motivational quote ("You don't have to struggle in silence..."), aiming to validate the user's struggle immediately upon opening the conceptual application.

Upon successful entry, the user is presented with the main dashboard shown in Fig. 7. The interface is designed for simplicity, with a focus on readability, making it easy for users to follow the navigation. To minimize cognitive load during moments of distress, the proposed dashboard centers on a single, prominent "Start Chat Now" button, accompanied by a personalized greeting to build a sense of companionship. This design enables users to interact naturally and initiate support sessions without navigating complex menus.

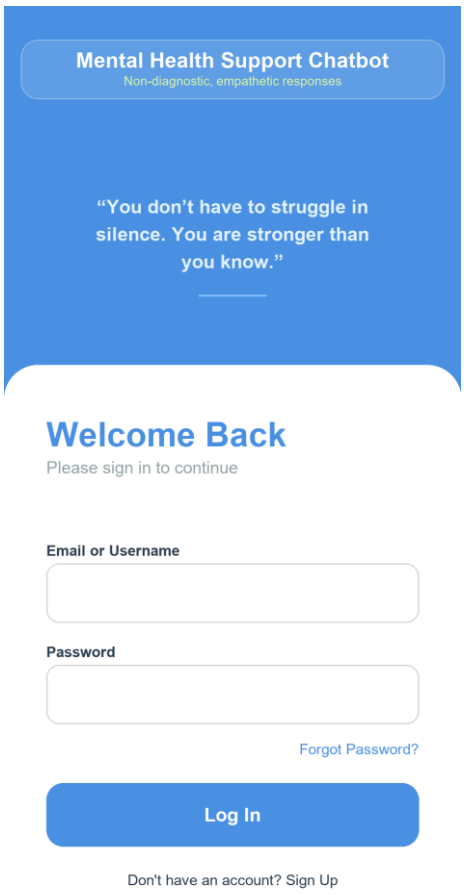

Figure 6. Login Interface with Motivational Quote

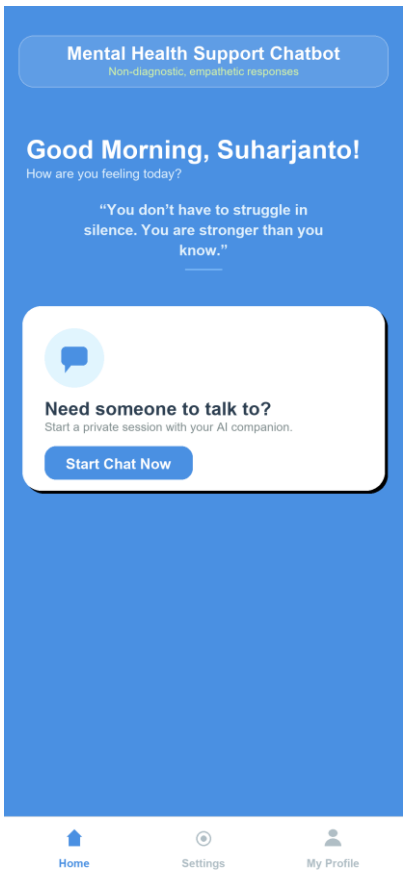

Figure 7. Main Dashboard and Navigation Menu

The structural logic governing the user interaction is illustrated in the system flowchart in Fig. 8. The interaction begins when the user opens the application and proceeds through a structured flow from the dashboard to the conversation interface. A critical component of this flow is the automated decision node for "Crisis Risk Detected", which determines whether the system should generate a standard empathetic response or trigger a crisis intervention protocol. This ensures that the interaction proceeds iteratively while maintaining safety guardrails.

Before the conversation begins, the user is presented with brief information, including a disclaimer and the ethical limitations of chatbot use, as shown in Fig. 9. Upon agreement by tapping "I Understand & Start Chat", the user is directed to the main conversation interface. This step ensures users are fully aware that the system is a supportive tool and not a substitute for professional medical diagnosis.

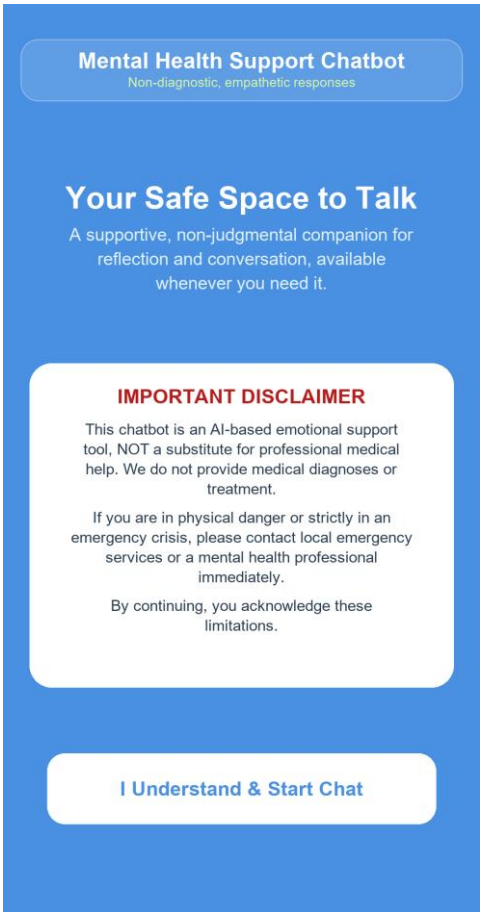

Figure 9. Ethical Disclaimer and Landing Page

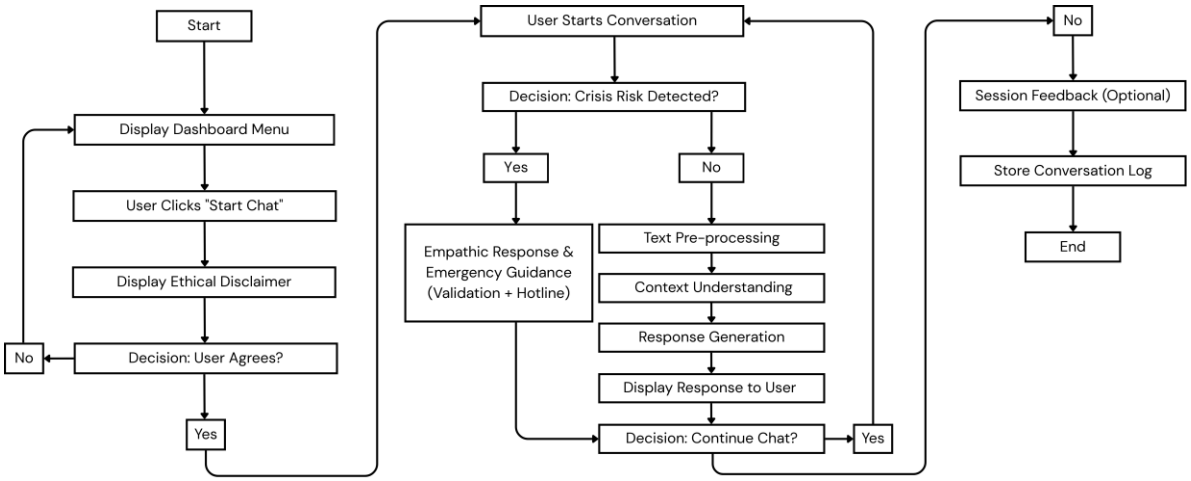

Figure 8. User Interaction and System Flowchart

Once the session begins, Fig. 10 displays the conceptual chatbot interface design for non-crisis conditions. The interface consists of a chat area displaying the dialogue between the user and the chatbot, a text input field, and a message send button. Every chatbot response is presented in clear, empathetic language, avoiding technical jargon and diagnostic claims. In this figure, the system processes user input about feelings of emptiness and generates a response focused on emotional support and reflection, aligned with the conversation context.

However, if risk indications or crisis conditions are detected during the conversation, the system specifically provides empathetic responses accompanied by professional help information. Fig. 11 demonstrates this critical functionality. When the system detects high-risk keywords (e.g., "I can't take it anymore"), the interface is designed to shift to a visual warning state and provide direct recommendations to contact suicide prevention services or relevant hotlines. During the session, the system automatically maintains conversation context, allowing interaction to proceed iteratively without requiring explicit actions to continue. Users possess full control to continue the conversation by sending follow-up messages or to end the session at any time. This proposed flow is designed to be intuitive and structured, resembling digital health application interaction patterns, with the aim of maintaining user comfort and minimizing cognitive load.

4 CONCLUSION

This study successfully developed and evaluated a mental health chatbot system based on the FLAN-T5-base architecture, designed to provide accessible initial emotional support. Through comparative analysis, the Full Fine-Tuning (FFT) approach yielded slightly superior semantic coherence and emotional tailoring. However, the Low-Rank Adaptation (LoRA) approach demonstrated remarkable computational efficiency by optimizing only 884,736 trainable parameters (~0.35%). This strategy reduced VRAM usage by approximately 21% (from 14.2 GB to 11.2 GB) and reduced the training time by approximately 24%. Crucially, traditional n-gram metrics such as BLEU-4 and ROUGE yielded objectively low scores, highlighting their limitations in evaluating open-ended empathetic dialogue. However, the subsequent LLM-as-a-judge evaluation empirically confirmed the system's true viability. Both FFT and LoRA achieved near-perfect scores in ethical safety (4.91 and 4.83 out of 5, respectively) and demonstrated a commendable baseline in empathetic validation, proving that the parameter-efficient approach produces safe and relevant support rather than nonsensical text.

This study used the Eco-Socio-Psycho-Religio-Technic (ESPRT) concept as a post-hoc analytical framework. This approach provides a multidimensional perspective for synthesizing and evaluating the essential aspects a conversational system in the mental health domain must possess. Viewed through this holistic lens, the LoRA-based system offers a pragmatic balance. It provides a sustainable, resource-efficient solution (Eco) that lowers the technical barrier for widespread social accessibility (Socio). Psychologically (Psycho), the qualitative evaluation confirms it maintains sufficient empathy for initial support despite its structural brevity. The system strictly adheres to ethical safety standards via a non-diagnostic design (Religio), and its high parameter efficiency makes it technically feasible for development and future integration in constrained hardware environments (Technic).

Future research should explore transitioning this system from a conceptual mockup into a fully deployed, real-world application. Integrating the fine-tuned model into a live platform and conducting direct user testing will be essential to evaluate its practical utility, real-time latency, and overall user comfort in actual mental health support scenarios.

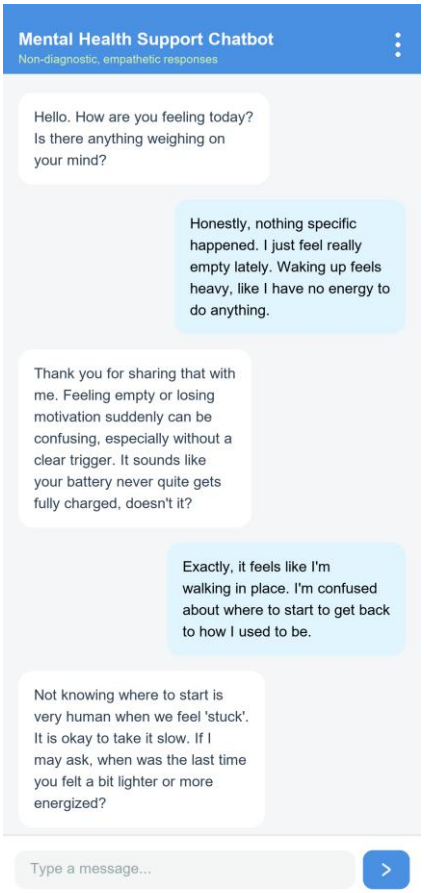

Figure 10. Normal Empathetic Conversation Interface

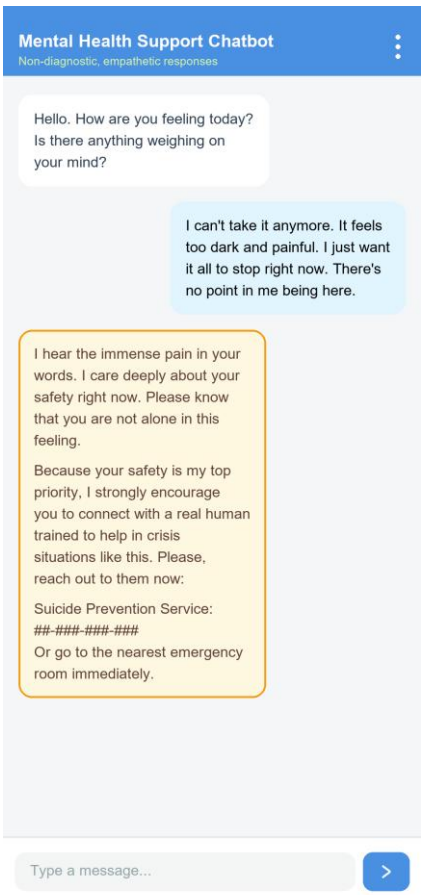

Figure 11. Crisis Response and Safety Guardrail Interface

CREDIT AUTHOR STATEMENT

Author 1 led the research design, experiment, analysis, and manuscript preparation. Authors 2 and 3 contributed to the interpretation of the findings.

COMPETING INTERESTS

The authors declare no competing financial or non-financial interests. The study was conducted independently, with no external funding sources influencing the research design, data analysis, or the decision to publish.

DECLARATION OF GENERATIVE AI AND AI-ASSISTED TECHNOLOGIES IN THE WRITING PROCESS

The authors acknowledge the use of generative AI tools exclusively for linguistic refinement, grammatical correction, and formatting purposes during manuscript preparation. The authors retain full responsibility for the manuscript's intellectual content.

ACKNOWLEDGMENT

We express our gratitude to Xu et al. [31] for making the MentalChat16K dataset publicly available for research purposes.

REFERENCES

[1] A. Magomedova and G. Fatima, "Mental Health and Well-Being in the Modern Era: A Comprehensive Review of Challenges and Interventions," *Cureus*, vol. 17, no. 1, p. e77683, 2025, doi: 10.7759/cureus.77683.

[2] M. Moitra *et al.*, "Global Mental Health: Where We Are and Where We Are Going," *Curr. Psychiatry Rep.*, vol. 25, no. 7, pp. 301–311, 2023, doi: 10.1007/s11920-023-01426-8.

[3] World Health Organization, "Mental Health: Strengthening our Response," Oct. 2025. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/mental-health-strengthening-our-response

[4] W. Liu *et al.*, "Global burden and trends of major mental disorders in individuals under 24 years of age from 1990 to 2021, with projections to 2050: insights from the Global Burden of Disease Study 2021," *Front. Public Health*, vol. 13, 2025, doi: 10.3389/fpubh.2025.1635801.

[5] J. Xu *et al.*, "The burden of mental disorders, substance use disorders, and self-harm among youths globally: findings from the 2021 Global Burden of Disease study," *Transl. Psychiatry*, vol. 15, no. 1, p. 346, 2025, doi: 10.1038/s41398-025-03533-x.

[6] L. Munira, P. Liamputtong, and P. Viwattanakulvanid, "Barriers and facilitators to access mental health services among people with mental disorders in Indonesia: A qualitative study," *Belitung Nurs. J.*, vol. 9, no. 2, pp. 110–117, 2023, doi: 10.33546/bnj.2521.

[7] Y. Gao, R. Burns, L. Leach, M. R. Chilver, and P. Butterworth, "Examining the mental health services among people with mental disorders: a literature review," *BMC Psychiatry*, vol. 24, no. 1, p. 568, 2024, doi: 10.1186/s12888-024-05965-z.

[8] M. Njuguna, I. Ngune, and Y. Middlewick, "Mental Health Support Workers Perspectives on Barriers to and Facilitators of the Effective Delivery of their Roles: A Systematic Review and Meta-aggregation," *Community Ment. Health J.*, vol. 61, no. 8, pp. 1548–1573, 2025, doi: 10.1007/s10597-025-01490-9.

[9] E. Mayor, "Chatbots and mental health: a scoping review of reviews," *Current Psychology*, vol. 44, no. 15, pp. 13619–13640, 2025, doi: 10.1007/s12144-025-08094-2.

[10] H. R. Lawrence, R. A. Schneider, S. B. Rubin, M. J. Matarić, D. J. McDuff, and M. J. Bell, "The Opportunities and Risks of Large Language Models in Mental Health," *JMIR Ment Health*, vol. 11, p. e59479, Jul. 2024, doi: 10.2196/59479.

[11] A. Bucher, S. Egger, I. Vashkite, W. Wu, and G. Schwabe, "'It's Not Only Attention We Need': Systematic Review of Large Language Models in Mental Health Care," *JMIR Ment Health*, vol. 12, p. e78410, 2025, doi: 10.2196/78410.

[12] Z. Guo, A. Lai, J. H. Thygesen, J. Farrington, T. Keen, and K. Li, "Large Language Models for Mental Health Applications: Systematic Review," *JMIR Ment Health*, vol. 11, p. e57400, 2024, doi: 10.2196/57400.

[13] R. Sanjeewa, R. Iyer, P. Apputhurai, N. Wickramasinghe, and D. Meyer, "Empathic Conversational Agent Platform Designs and Their Evaluation in the Context of Mental Health: Systematic Review," *JMIR Ment Health*, vol. 11, p. e58974, 2024, doi: 10.2196/58974.

[14] A. A. Dzaky *et al.*, "Optimization Chatbot Services Based on DNN-Bert for Mental Health of University Students," *Journal of Applied Informatics and Computing*, vol. 8, no. 1, pp. 13–21, 2024, doi: 10.30871/jaic.v8i1.7403.

[15] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *CoRR*, vol. abs/1810.04805, 2018, [Online]. Available: http://arxiv.org/abs/1810.04805

[16] J. Zeniarja *et al.*, "Integrating ELECTRA and BERT models in transformer-based mental healthcare chatbot," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 37,

no. 1, pp. 315 – 324, 2025, doi: 10.11591/ijeecs.v37.i1.pp315-324.

[17] K. Clark, M.-T. Luong, Q. V Le, and C. D. Manning, "ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators," *CoRR*, vol. abs/2003.10555, 2020, [Online]. Available: https://arxiv.org/abs/2003.10555

[18] D. Mandal and H. Himanshu, "Harnessing Deep Learning for Mental Health: A Comparative Study of BERT, LSTM, and GPT-2," in *2025 6th International Conference for Emerging Technology (INCET)*, 2025, pp. 1–7. doi: 10.1109/INCET64471.2025.11140291.

[19] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language Models are Unsupervised Multitask Learners," *OpenAI*, 2019, [Online]. Available: https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

[20] J. S. J, K. Manoharan, and K. Arul, "Evaluation of BERT and BART in Mental Health Dialogue by Using a Counselling Chatbot," in *2025 International Conference on Engineering Innovations and Technologies (ICoEIT)*, 2025, pp. 1412–1417. doi: 10.1109/ICoEIT63558.2025.11211681.

[21] M. Lewis *et al.*, "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds., Online: Association for Computational Linguistics, Jul. 2020, pp. 7871–7880. doi: 10.18653/v1/2020.acl-main.703.

[22] O. Taiwo and B. Al-Bander, "Emotion-aware psychological first aid: Integrating BERT-based emotional distress detection with Psychological First Aid-Generative Pre-Trained Transformer chatbot for mental health support," *Cognitive Computation and Systems*, vol. 7, no. 1, p. e12116, 2025, doi: 10.1049/ccs2.12116.

[23] H. Q. Yu and S. McGuinness, "An Experimental Study of Integrating Fine-tuned LLMs and Prompts for Enhancing Mental Health Support Chatbot System," *J. Med. Artif. Intell.*, vol. 7, p. 16, 2024, doi: 10.21037/jmai-23-136.

[24] H. M. Mohssen and H. H. Safi, "An Efficient Mental Healthcare Chatbot Using Reprogramming Transformer Model Based on a Fine-Tuned GPT-2," in *Data Processing and Networking*, A. Swaroop, B. Virdee, S. D. Correia, and J. Valicek, Eds., Singapore: Springer Nature Singapore, 2025, pp. 69–83.

[25] L. Wang *et al.*, "Parameter-efficient fine-tuning in large language models: a survey of methodologies," *Artif. Intell. Rev.*, vol. 58, no. 8, p. 227, 2025, doi: 10.1007/s10462-025-11236-4.

[26] Z. Han, C. Gao, J. Liu, J. Zhang, and S. Q. Zhang, "Parameter-Efficient Fine-Tuning for Large Models: A Comprehensive Survey," 2024. [Online]. Available: https://arxiv.org/abs/2403.14608

[27] E. J. Hu *et al.*, "LoRA: Low-Rank Adaptation of Large Language Models," *CoRR*, vol. abs/2106.09685, 2021, [Online]. Available: https://arxiv.org/abs/2106.09685

[28] V. B. Parthasarathy, A. Zafar, A. Khan, and A. Shahid, "The Ultimate Guide to Fine-Tuning LLMs from Basics to Breakthroughs: An Exhaustive Review of Technologies, Research, Best Practices, Applied Research Challenges and Opportunities," 2024. [Online]. Available: https://arxiv.org/abs/2408.13296

[29] V.-T. Doan, Q.-T. Truong, D.-V. Nguyen, V.-T. Nguyen, and T.-N. N. Luu, "Efficient Finetuning Large Language Models For Vietnamese Chatbot," in *2023 International Conference on Multimedia Analysis and Pattern Recognition (MAPR)*, 2023, pp. 1–6. doi: 10.1109/MAPR59823.2023.10288647.

[30] S. Pratap, A. R. Aranha, D. Kumar, G. Malhotra, A. P. N. Iyer, and S. S.S., "The fine art of fine-tuning: A structured review of advanced LLM fine-tuning techniques," *Natural Language Processing Journal*, vol. 11, p. 100144, 2025, doi: 10.1016/j.nlp.2025.100144.

[31] J. Xu *et al.*, "MentalChat16K: A Benchmark Dataset for Conversational Mental Health Assistance," in *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, in KDD '25. New York, NY, USA: Association for Computing Machinery, 2025, pp. 5367–5378. doi: 10.1145/3711896.3737393.

[32] H. W. Chung *et al.*, "Scaling instruction-finetuned language models," *J. Mach. Learn. Res.*, vol. 25, no. 1, Jan. 2024.

[33] C. Raffel *et al.*, "Exploring the limits of transfer learning with a unified text-to-text transformer," *J. Mach. Learn. Res.*, vol. 21, no. 1, Jan. 2020.

[34] Md. S. Salim, S. I. Hossain, T. Jalal, D. K. Bose, and M. J. I. Basher, "LLM based QA chatbot builder: A generative AI-based chatbot builder for question answering," *SoftwareX*, vol. 29, p. 102029, 2025, doi: 10.1016/j.softx.2024.102029.

[35] D. J. P and S. B. Karnati, "BuddyBot: AI Powered Chatbot for Enhancing English Language Learning," in *2024 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI)*, 2024, pp. 1–6. doi: 10.1109/IATMSI60426.2024.10502595.

[36] B. B. P. Sai, R. H. Krishna, C. L. Reddy, and C. U. Kumari, "Enhancing Conversational AI: Building an Intelligent Chatbot with Llama Using LoRa & Analyzing Different Llama Models," in *2025 International Conference on Inventive Computation Technologies (ICICT)*, 2025, pp. 1061–1066. doi: 10.1109/ICICT64420.2025.11004777.

[37] Y. Mao *et al.*, "A survey on LoRA of large language models," *Front. Comput. Sci.*, vol. 19, no. 7, p. 197605, 2024, doi: 10.1007/s11704-024-40663-9.

[38] J. Hu, X. Liao, J. Gao, Z. Qi, H. Zheng, and C. Wang, "Optimizing Large Language Models with an Enhanced LoRA Fine-Tuning Algorithm for Efficiency and Robustness in NLP Tasks," in *2024 4th International Conference on Communication Technology and Information Technology (ICCTIT)*, 2024, pp. 526–530. doi: 10.1109/ICCTIT64404.2024.10928552.