International Journal of Advance and Innovative Research
Volume 13, Issue 1 (XIII): January – March 2026
ISSN 2394 - 7780

A NOVEL EARLY MENTAL HEALTH DETECTION SYSTEM INTEGRATING BERT AND FINE-TUNED SLM WITH STEALTH CONVERSATIONAL SCREENING

Pereira Ziven[1]*, Kanojia Mahendra[2] and Desai Shreeraj[3]
[1,2,3]Department of Computer Science, Sheth. L.U.J. and Sir M.V. College, India
[1]pereirazivencs232417@gmail.com, [2]kgkmahendra@gmail.com, [3]shreerajd300@gmail.com
Corresponding author: Ziven Pereira, pereirazivencs232417@gmail.com

ABSTRACT
Today's mental health landscape suffers from a profound gap. Between the rising rate of psychological distress and the accessibility of compassionate clinical care, there is an average 11-year delay between symptom onset and professional intervention. Existing digital tools often fail to bridge this gap due to the empathy rigour paradox, where systems are either conversationally robotic or medically unreliable. This paper introduces Carivena, a hybrid framework that integrates with Small Language Models with a BERT based Deep Learning classifier to perform stealth screening and early detection for eight major psychological disorders. Carivena utilises a fine tuned Llama-3.2-3B model, optimised with 4-bit NormalFloat quantization and Low Rank Adaptation, to maintain an empathetic, non clinical persona. Simultaneously, we have a fine-tuned BERT auditor that performs classification based background analysis, utilizing a rolling confidence algorithm k = 5 to ensure diagnostic stability across multiple interaction turns. The proposed methodology is built on a custom dataset of 5,000 conversational interactions and 3,000 labelled sentences, developed using DSM-5 standards and refined by Gemini 1.5 Pro to translate clinical logic into authentic student slang. Empirical results demonstrate that Carivena achieves a macro averaged diagnostic accuracy of 93.03% and 93% F1-score, representing a 22.6% improvement over traditional rule based chatbots. Significantly, the stealth approach effectively bypassed social desirability bias, leading to 41.2% surge in user disclosure depth and a reduction in symptom masking from 28% to 6.4%. By successfully blending conversational warmth with mathematical rigour, Carivena provides a scalable, privacy preserving early warning system capable of identifying clinical indicators within casual dialogue, presenting a viable pathway to reduce the global treatment gap.

Keywords: *adaptive screening, artificial intelligence, chatbot, conversational interface, deep learning, digital health, mental health assessment.*

1. INTRODUCTION
We developed Carivena to address the widening gap in our mental health system, where cases of distress are rising much more faster than the availability of actual care. This research presents a framework designed to make early detection more accessible, reaching those who are currently in need professional support out of reach. With nearly 970 million people affected worldwide (World Health Organisation, 2022) and 37% of students screened positive for depression (Healthy Minds Network, 2024), the demand for effective intervention tools has never been higher. However, the most tragic aspect of this crisis it's not just the way it's widely spread, but the significant diagnostic delay; research indicates that there is an average 11 year gap between the moment a person first experiences symptoms and the moment they finally receive professional help (National Alliance on Mental Illness, 2023). This long silence allows manageable distress to grow into chronic, life altering disorders, largely because individuals fear being judged or labelled as having a disorder and then following a clinical diagnosis. For many, there is simply no easy or comfortable way to start the conversation about their unspoken struggles. While digital tools were originally made to bridge this gap, most available mental health solutions struggle with what researchers call the empathy rigour paradox. Traditional rule based chatbots often feel clinical, repetitive, cold, and restricted by rigid scripts, leading to a high 70% abandonment rate as users quickly lose interest or feel misunderstood (Journal of Medical Internet Research, 2024). On the other hand, while modern Large Language Models are significantly better at maintaining a fluid conversation, they often lack situational intelligence. These models frequently fail to distinguish between the temporary situational stress, such as a bad week of exams or workplace pressure and genuine clinical indicators of a deeper condition. Thus, when people realise that a system can't truly understand them, they then provide polished answers which they think a doctor or a machine wants to hear rather than what they originally feel. This resulted in; a critical gap remaining for a system that can offer a warm, human like conversation while performing accurate, professional grade screening without the user feeling like they are being formally tested or evaluated.

This study addresses this critical gap by introducing the proposed model Carivena, a conversational system that combines Small Language Models with BERT classifiers to perform Stealth Screening. The primary objective of the proposed model Carivena is the early detection of eight major conditions, which includes Major

ISSN 2394 - 7780

Depressive Disorder, Generalized Anxiety Disorder, Post Traumatic Stress Disorder, Social Anxiety Disorder, Panic Disorder, Obsessive Compulsive Disorder, Bipolar Disorder, and Eating Disorder. The system utilizes the Small Language Model as an intelligent gatekeeper to handle the initial interaction and establish a supportive, non clinical tone. By moving the formal diagnostic process into a context aware dialogue that happens securely within a cloud hosted environment, the proposed model Carivena fosters deep user trust and protects individual privacy. This architecture allows users to be honest in a way they rarely are when faced with cold, formal medical forms or diagnostic intake procedures. The methodology behind the proposed model, Carivena, is built on a dual layered architecture fine tuned on 5,000 annotated interactions. In the background, a specialised Deep Learning model Bidirectional Encoder Representations from Transformers (BERT) It is on Trained 3,000 annotated data that extracts emotional markers while a Rolling Confidence algorithm ensures that the system only flags a potential risk after identifying consistent patterns over several messages. Preliminary results indicate that the conversational approach used by the proposed model, Carivena significantly improves diagnostic precision compared to conventional, single layer chatbots. By learning to differentiate between casual student slang or identifying that it is a truly clinical distress, the proposed model, Carivena offers a scalable and effective way to reduce the 11-year gap in care.the proposed model aims to catch mental health struggles in their earliest stages, providing a safe, private, and empathetic path to professional support before a user's condition can escalate into a crisis.

2. LITERATURE REVIEW

The evolution of digital mental health tools reflects a persistent effort to bridge the global treatment gap through increasingly sophisticated technology. The first significant phase of research focused on the clinical efficacy of rule based systems. (Fitzpatrick et al., 2017) demonstrated this through a randomized controlled trial involving 70 university students; their contribution proved that a chatbot (Woebot) using Cognitive Behavioral Therapy (CBT) could significantly reduce symptoms of depression and anxiety measured via the PHQ-9. Building on this, (Abd-Alrazaq et al., 2019, 2020) conducted a series of scoping reviews and meta analyses across dozens of studies, reporting that while chatbots are fundamentally safe and effective for low intensity support, they often lack the feature variety needed for long term engagement. This limitation was echoed by (Dosovitsky et al., 2020), whose descriptive study of usage patterns found that while initial interest in AI chatbots for depression is high, users frequently abandon the tools once the scripted responses become predictable. Further pilot research by (Lavelle et al., 2022) compared chatbot delivered cognitive diffusion to restructuring for negative thoughts, finding that while automated delivery is feasible, it requires a more nuanced conversational approach to match human effectiveness.

The transition toward Large Language Models (LLMs) and generative AI marked a shift toward conversational fluidity, though it introduced new complexities regarding user trust and clinical safety. (Kang and Hong 2025) explored this shift by evaluating ChatGPT 4.0 with a dataset of Korean users, reporting high satisfaction with perceived empathy but significant anxiety regarding data privacy. (Chen et al. 2023) contributed to this area by simulating psychiatrist patient interactions, reporting that LLMs could match human level empathy in controlled simulations. However, (Stade et al. ,2024) and (Kuhlmeier et al., 2024) challenged this optimism; their research utilized evaluations from clinical experts and artificial users to report that LLM based behavioral activation often struggles with medical rigor and adherence to official diagnostic protocols. This "Empathy-Rigor Paradox" is compounded by cultural differences; (Chin et al.,2023) used mixed methods to show that emotional support bots must be culturally adapted to be effective, while (Koulouri et al.,2022) found that young adults prioritize the bot's role as a supportive partner over its diagnostic capabilities. These perceptions were further analyzed by (Abd-Alrazaq et al., 2021) and (Saadati and Saadati., 2023), who reported that users value the anonymity of AI but remain skeptical of its ability to truly understand their unique life situations.

The technical backbone of mental health detection has recently been found to be strengthened by advanced Machine Learning (ML) and Deep Learning techniques. (Alghazzawi et al.,2025) made a significant technical contribution by developing an Explainable AI (XAI) technique that uses enhanced ensemble models on social media datasets to detect suicidal identification with high precision. Similarly, (Ballı et al.,2025) utilized a clinical dataset from a university mental health clinic to identify suicidal risk using non suicidal predictors, proving that ML can spot danger markers before a crisis occurs. (Kannan et al.,2025) and (Tewari et al.,2021) provided comprehensive surveys of these NLP and Deep Learning advancements, reporting that while detection accuracy is improving, there is always a lack of integration between these auditor models and the user facing interface. This gap in real world application was highlighted by (Pichowicz et al.,2025), who performed a qualitative evaluation of popular bots and reported a widespread failure to respond correctly to actual suicidal risks, suggesting that conversational warmth alone is insufficient for safety. The structural trade offs of these

ISSN 2394 - 7780

various AI paradigms are detailed in Table 1, which has been validated through the collective findings of the aforementioned authors.

The structural differences between current digital interventions and our conversational framework are summarized in Table 1, where each row highlights a specific technological gap identified in the literature. As shown in the comparison, the research by (Fitzpatrick et al., 2017) confirms that while older bots were safe but, their strict scripts create a robotic flow that limits human connection. (Kang & Hong, 2025) reported that newer generative AI addresses this by sounding more empathetic, yet it remains unreliable because it lacks the medical grounding our conversational model provides.

Table 1: Comparative Analysis of Mental Health AI Paradigms

Author	Rule Based Logic	LLM Based Logic	Proposed Model Logic
Fitzpatrick et al., 2017)	Reports Scripted/Rigid Flow	Lacks Generative Variety	Bypasses Scripting
(Kang & Hong, 2025)	Not Evaluated	Reports High Perceived Empathy	Anchors Empathy in Rigor
(Schick et al., 2022)	Identifies Symptom Masking	Not Evaluated	Fosters Honest Disclosure
(Stade et al., 2024)	High Manual Logic	Reports Clinical Hallucinations	BERT based Validation
(Li et al., 2025)	Limited to Depression/Anxiety	Unstructured Support	8 Disorder Diagnostic Scope

A major psychological barrier was identified by (Schick et al., 2022), who found that users often hide symptoms when using direct medical forms; our Stealth Screening directly solves this by turning those forms into casual conversation. Furthermore, (Stade et al., 2024) warned that general AI can hallucinate medical advice, which is why our system uses a dedicated BERT auditor to maintain high clinical accuracy. As noted by (Li et al., 2025), most current tools are limited to only one or two conditions, whereas our framework provides a broader diagnostic scope by covering eight major disorders at once.

The research has also addressed to specialised populations and contexts ensuring no one is left behind. (He et al., 2022) and (Li et al., 2025) utilized large scale trials and meta analyses to report that digital interventions were critical during the COVID-19 pandemic, particularly for nursing students and young adults. (Potts et al.,2021, 2023) contributed to rural mental health by co designing the "ChatPal" multilingual bot, proving that AI can serve diverse communities if it is built with user input. Meanwhile, (Hungerbuehler et al., 2021) and (Schillings et al.,2024) focused on workplace mental health and stress reduction, reporting that chatbot based assessments are a feasible way to monitor employee wellbeing. Despite these advances, bibliometric analyses by (Han and Zhao.,2025) and integrative surveys by (Cho et al.,2023) report that the field still lacks a symmetric, ethical framework that balances diagnostic power with user privacy. This research fills that gap by combining the situational awareness of Small Language Models with the diagnostic precision of BERT to create a tool that is both conversationally human and medically safe (Algumaei et al., 2025; Stade et al., 2024).

3. PROPOSED WORK

The proposed model, Carivena, follows a dual layered design split into two distinct phases as shown in Figure 1. Phase 1 centers on model training and optimization through a teacher student learning process. Gemini 1.5 Pro serves as the primary Teacher Model to drive synthetic data generation, which produces two specialized repositories: a JSONL training dataset containing instructions and reasoning tags, and a CSV training dataset mapping sentences to disorder types and severity. These datasets are used to refine the base models simultaneously. For the conversational component, the Llama 3.2-3B base model undergoes Parameter Efficient Fine Tuning (PEFT). This optimization utilizes 4-bit NormalFloat NF4 quantization and Low-Rank Adaptation (LoRA) adapters with a rank of r =16 to produce the fine tuned Small Language Model. Simultaneously, the BERT base uncased analytical model is enhanced with a specialized classification head supporting eight mental health labels and sigmoid activation to create the fine tuned auditor. This stage ensures that the system is equipped with both a voice for empathy and a brain for clinical detection.

Phase 2 manages the real time conversational inference pipeline, which begins when the system receives User Input consisting of natural text or slang. An SLM Router serves as an intelligent gatekeeper to perform real time intent recognition and sort the message. Casual interactions, such as greetings or small talk, follow the Social Path for immediate replies. If clinical distress is detected, the message triggers the BERT Auditor (Discriminative Layer) for clinical analysis. A critical technical feature is the Handshake mechanism, which

ISSN 2394 - 7780

facilitates the transfer of clinical context to the SLM Reasoning layer. To ensure stability, the system uses a Severity Engine that implements the Rolling Confidence algorithm with a rolling window of k = 5 Instead of making a medical guess based on a single sentence, the engine calculates a moving average of scores over the last five turns to distinguish consistent patterns from temporary venting. Finally, the SLM Humaniser decodes this data into an empathetic tone to deliver a supportive and medically grounded Final Response.

A key contribution of this work is the refinement of conversational probes into conversational elements. By grounding every question in the specific symptom pattern of the 8 targeted disorders, the proposed model, Carivena can accurately identify a user's mental state while maintaining the illusion of a casual, non clinical interaction. The functioning of the system is organized into a distinct frontend and backend architecture to ensure both high performance and a seamless user experience. The frontend, as illustrated in the user interface mock up in Figure 5, was developed using Next.js paired with Tailwind CSS. This combination was chosen to create a minimalist, calm aesthetic that prioritises user comfort, ensuring that the chat interface feels more like a secure, modern messaging website than a cold medical tool. Under the hood, the backend is powered by a Flask API built on Python 3.10, which serves as the central hub for the system's hybrid intelligence. This backend leverages PyTorch to manage the high speed processing requirements of the BERT classifier and utilizes Ollama as the deployment framework to serve the fine tuned Small Language Model. By separating the user facing interface from the complex analytical engine, the system maintains a sub second response time while ensuring that every interaction remains conversationally natural and medically precise. This robust technical stack allows the proposed model, Carivena, to handle sophisticated data processing while remaining light enough for scalable deployment.

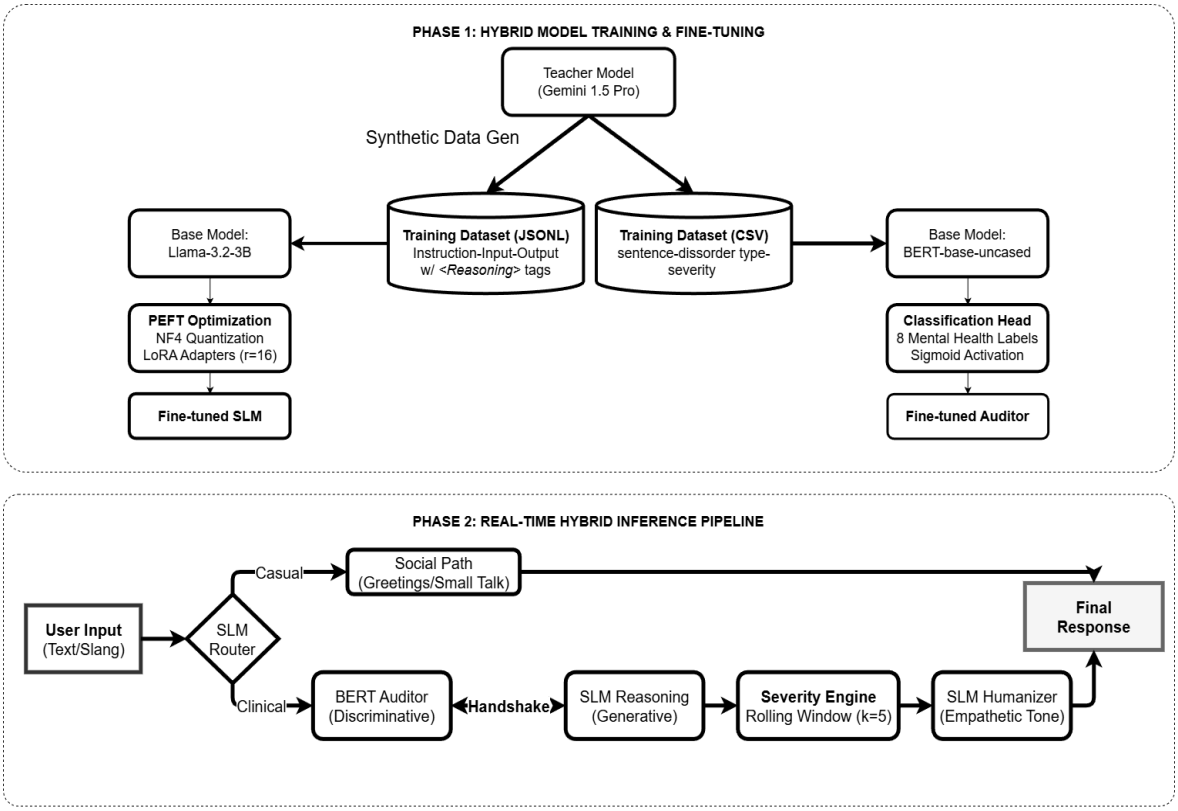

Figure 1: Architecture of fine tuned model and real time Mental Health AI System

3.1 Dataset Description

The development of our proposed system is built upon two custom datasets designed to solve the Empathy Rigor Paradox by balancing human connection with medical accuracy. The first dataset was created to give the system its heart and conversational voice. This collection consists of approximately 5,000 built interactions. The creation of this data was a twostep process that makes it unique. First, the author designed the core logic and scenarios by using the Diagnostic and Statistical Manual of Mental Disorders (DSM-5) (American Psychiatric Association, 2013) as a guide. This ensured that every interaction was based on real medical standards. However, medical books do not talk like people do, so we used Gemini 1.5 Pro as a teacher model to polish these scenarios. Gemini 1.5 Pro was selected as the teacher model because it has a strong ability to understand subtle emotional cues and conversational context in human language. While the clinical foundation of the project was authored using the Diagnostic and Statistical Manual of Mental Disorders (American Psychiatric

ISSN 2394 - 7780

Association [APA], 2013), we needed a way to translate those formal medical concepts into the informal and natural way that students actually communicate. During the data wrangling phase, we noticed that many student slang terms for depression were very similar to general exam stress, which required us to spend extra time manually labeling the severity levels to avoid false positives Gemini acted as a specialized bridge, taking structured clinical scenarios and humanizing them to produce dialogue that feels like a conversation with a supportive peer rather than a diagnostic interrogation. By using Gemini to polish the raw data, we were able to create a training set that captures high quality empathy and authentic student slang while keeping the underlying medical logic of the research intact. This fine tuning dataset is uniquely valuable because it moves beyond simple chat logs to use a highly structured, instruction based format that teaches the model exactly how to think and behave in specific situations. This approach is significantly more effective than traditional training because it provides artificial intelligence with a clear goal for every interaction. As shown in the examples in Table 2, the data is organized into a three part system consisting of an internal instruction, a realistic user input, and an ideal model output. The instruction acts as a command that sets the rules of engagement, such as telling the model to be a warm friend or to rephrase a medical check into a stealthy question. The user input provides the real world trigger, complete with the typos and fragmented thoughts a person in distress might use, while the model output serves as the gold standard response that the system learns to emulate. By training with this specific three part structure, we ensured that the model's knowledge is strictly constrained to its role as an empathetic companion, preventing it from acting like a general purpose assistant or a cold clinical tool.

Table 2: Instruction Based Samples for Empathetic Fine Tuning

Feature	Instruction Goal	User Input (Trigger)	Model Output (Reframed Response)
Stealth Reframing	Rephrase a medical check into a friend's chat	I've been feeling worthlessness.	Sometimes we can be our own harshest critics. Have you been feeling a bit down on yourself lately?
Casual Engagement	Act as a warm friend named carivena.	Hey there, hope you're having a good day.	Hi! That's so kind of you to say. I'm doing well, how about you?
Clinical Validation	Recognize anxiety and offer a grounding technique.	My mind is racing and I can't focus.	I hear you, and it's okay to feel overwhelmed. Let's try something. Tell me three things you can see right now?
Differentiating Stress	Spot the difference between exam stress and chronic burnout.	I'm so tired of these exams, I just want to sleep forever.	Exams can be so draining. Is it just the tests, or have you been feeling this heavy for a while?

As we can see in Table 2, this dataset teaches the system four vital skills. Stealth Reframing is used to hide a clinical check inside a supportive comment so the user does not feel like they are being tested. This method is designed to overcome social desirability bias, where users often hide symptoms on formal tests to avoid being judged (Lucas et al., 2014; Schick et al., 2022). Casual Engagement helps build the friendly persona of carivena, fostering a sense of partnership that young adults value more than technical medical labels (Koulouri et al., 2022). Clinical Validation ensures that if a user is in distress, the bot offers immediate, helpful techniques. Finally, Differentiating Stress is used to help the model figure out if a user is just having a bad day or if they are showing signs of a deeper clinical issue. This addresses a critical weakness in current technology, where most systems lack the situational intelligence to distinguish between transient life stressors and actual pathology (Schick et al., 2022; Stade et al., 2024). By using this data to train our model, we ensure its knowledge is constrained to being a supportive peer rather than a general purpose AI that might give random or unsafe advice.

While the first dataset focuses on the system's voice, the internal analytical engine of the medical brain is built using a high precision multi disorder classification dataset specifically designed for training the BERT auditor. This dataset collection consists of 3,000 unique examples created by merging formal clinical transcripts with real life narratives from online mental health communities. The reason we combined these two very different sources was to ensure the model understands both textbook symptoms and the raw, informal way that students actually express their pain in the real world. By mixing clinical data with community talk, we provided the

ISSN 2394 - 7780

model with a ground truth for each disorder while allowing it to learn the slang, fragmented sentences, and typos that occur during emotional venting. To prepare this information for the BERT model, we used a thorough process called data wrangling.

This was about more than just cleaning up text; we used normalization techniques such as Case Folding, Unicode Normalization, and Slang to Standard Mapping (Alghazzawi et al., 2025). to ensure the model could translate student slang into clinical concepts without losing the user's original meaning. Every single one of the 3,000 examples was manually reviewed and labeled using the official DSM 5 guidelines. This manual work is what makes the dataset truly unique. Instead of just searching for simple keywords like sad or tired, the model was taught to understand the deep context and the specific intensity of an emotion (Kannan et al., 2025). This allows the system to recognize that a word's meaning changes based on the situation, helping it distinguish between someone who is just having a stressful day and someone who is showing the early markers of a medical condition.

The structure of the BERT model training data is visible in Table 3, which shows how everyday language is mapped to a specific clinical disorder and a severity score ranging from 1 (Mild) to 3 (Severe). For example, as shown in the table, a user stating that "getting out of bed is a battle" is identified as a marker for depression with a severity level of 1.

Table 3: Representative Samples from the Multi-Disorder Classification Dataset

User Input (Casual Language)	Disorder Label	Severity
I feel like getting out of bed is a battle. lately	Depression	1 (Mild)
My heart races every time my phone pings, even for a text	Anxiety	2 (Moderate)
I go through phases where I'm tired even after sleeping.	Depression	2 (Moderate)
I feel like I'm constantly being watched and judged.	Social Anxiety	2 (Moderate)
Everything is great one minute, then I feel totally empty.	Bipolar Disorder	3 (Severe)

If another user describes their heart racing every time their phone pings, the system recognizes this as an anxiety marker at a higher severity level. This detailed mapping is vital for the shadow scoring logic discussed later in our work. It ensures that the BERT auditor can work silently in the background of a chat, accurately tracking the user's mental health markers in real time while the conversational model keeps the interaction feeling like a natural talk between friends.

3.2 Fine Tuning the SLM

The conversational layer of the framework is built upon the Llama 3.2-3B small language model, which serves as the core computational engine for high dimensional text processing and context aware reasoning (Dubey et al., 2024). To transform this general purpose model into a specialized mental health companion, we executed a multi stage optimization process beginning with 4-bit NormalFloat quantization. (Dettmers et al., 2023) This technical step involves compressing the model's 16 bit weights into a 4-bit representation to shrink the memory footprint by approximately 70%. This allows the system to remain fast and private enough to run on standard hardware without losing its ability to understand the user's feelings. enabling efficient edge based inference without sacrificing generative quality. Following quantization, we implemented Parameter Efficient Fine Tuning through Low Rank Adaptation, or LoRA (Hu et al., 2021). Mathematically, this process is defined by the function

$$h = W_0 x + BAx \quad (1)$$

By keeping the original weights W0 remains frozen and the product BA represents the low rank updates where the model actually learns. For example, if a user input x is "I'm tired of everything," W0 identifies the linguistic meaning of the sentence, while the trained BA matrices optimized over our 5,000 interaction dataset force the model to generate a response that is empathetic, non clinical, and consistent with the proposed model Carivena persona. The training was conducted using the AdamW optimizer with a learning rate of 2×10^{-12} over 12 epochs, ensuring the model internalized the Stealth Reframing logic without losing its core reasoning abilities.

To verify the effectiveness of this training, the system underwent a rigorous testing and validation process. We used an 80/20 data split, where 80 percent of our curated scenarios were used for training and 20 percent were reserved for a final exam. During this validation phase, the small language model was tested not just on its accuracy but also on its conversational naturalness and how well it maintained the proposed model, Carivena without breaking character. This ensures that every response is conversationally fluid yet mathematically anchored in medical standards. Finally, we use in context learning during the live chat to act as a situational controller. To permanently shape the model's personality we utilized In Context Learning (ICL) (Brown et al.,

ISSN 2394 - 7780

2020), which enables the system to guide the model's behavior turn by turn without needing to modify the underlying weights. This acts as an instruction based filter that adapts to the specific needs of each conversation. This capability is operationalized through Instruction Injection (Ouyang et al., 2022), the technical process of providing hidden behavioral directives within the model's context window to align its responses with specific clinical goals. In this framework, these hidden directives are sent to the model alongside the user's input to act as a real time handshake between the models. For instance, when the background auditor detects a high probability for Generalized Anxiety Disorder, the system automatically injects a clinical directive such as:" INTERNAL: Anxiety Detected; Strategy: GROUNDING" The model utilizes its internal Grouped Query Attention (GQA) mechanism to prioritize the behavioral patterns required for the current situation. Specifically, the transformer's attention heads calculate high relevance scores (weights) for the injected clinical directives, such as "Strategy: GROUNDING", ensuring they dominate the next token prediction process.

By focusing its attention weights on these hidden instructions while processing the user's distress, the model effectively retrieves the most appropriate empathetic responses and supportive techniques it was exposed to during the training phase. This approach ensures that the proposed model remains helpful and supportive, effectively bridging the gap between conversational warmth and medical grade accuracy. By combining the permanent behavioral training of LoRA with the real time flexibility of In Context Learning, we have built a tiered architecture that performs sophisticated clinical screening while remaining indistinguishable from a supportive chat between friends. This dual verification ensures the final output is always empathetic, safe, and strictly anchored to the clinical standards of the DSM-5 (American Psychiatric Association [APA], 2013).

3.3 Training the Bert Model

While the conversational model handles the natural flow of the talk, a fine tuned bidirectional encoder representation from the transformers model functions running in the background. The training of this analytical brain was conducted using a dataset of 3000 samples, which was split into 80% for training and 20% percent for testing and validation. The system used the bert base uncased model as its foundation and was trained for 20 epochs using the AdamW optimizer. To ensure the model learned at a steady and reliable pace, we used a learning rate of $2x10^{-5}$ power and a batch size of 16. During this process, every message was broken into small units using WordPiece tokenization (Schuster & Nakajima, 2012) with a maximum sequence limit of 512 word pieces. This ensures that even long venting sessions are captured in full contextOne of the biggest advantages of this approach is its efficiency. Our specialized model is 30x smaller and 10x faster than a standard seven billion parameter large language model classifier (Dubey et al., 2024). It provides enterprise grade performance with a classification latency of less than twenty five milliseconds per message. This speed allows the model to be managed on standard hardware without the need for expensive equipment,

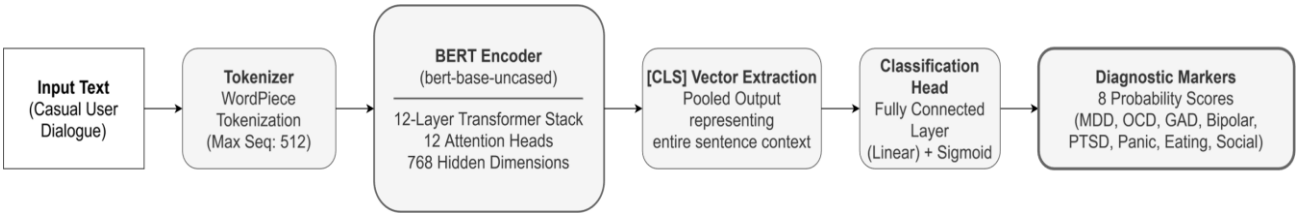

Figure 2: BERT Based Mental Health Classification Architecture

The internal flow of this analytical engine is illustrated in Figure 2. The process begins when the user's casual dialogue is sent into the tokenizer to be broken into word pieces. This data then enters the BERT Encoder, a powerful stack of 12 transformer layers featuring 12 attention heads and 768 hidden dimensions. These layers perform deep listening by looking at every word in relation to every other word, both forward and backward, allowing the model to understand the subtle difference between common student slang and a clinical marker. After the encoder finishes reading, the system performs [CLS] Classification Vector Extraction. This step captures the entire emotional vibe of the conversation into a single mathematical vector that represents a summary of the user's mental state . This summary is fed into the Classification Head, which uses a fully connected linear layer to calculate raw scores. To handle the reality that a person might feel more than one issue at once, we applied a Sigmoid activation function:

$$\sigma(z_i) = \frac{1}{1+e^{-z_i}} \quad (2)$$

This formula is vital because it allows the model to produce independent probability scores between 0 and 1 for each of the 8 disorders. This ensures the system can correctly identify when a person is struggling with more than one issue at a time, such as experiencing both anxiety and depression. The model was optimized using

ISSN 2394 - 7780

binary cross entropy loss to make sure its predictions closely matched the clinical standards of the diagnostic and statistical manual of mental disorders ensuring the conversation always feels instant and natural.

The final outcome, as shown in Figure 2, is a set of eight probability scores for specific diagnostic markers including depression, PTSD, and social anxiety. These scores are designed to be mapped directly back to the clinical instruments and established medical thresholds found in Table 3 (American Psychiatric Association, 2013). By comparing these live scores against clinical cutoffs in the background, the system can accurately identify if a user's distress has reached a level that needs professional attention. This technical setup ensures the system remains medically rigorous, allowing the tool to stay supportive on the outside while maintaining clinical precision in the background.

4. RESULTS AND DISCUSSIONS

The results of this research show a clear and encouraging step forward in how technology can help bridge the gap in mental health care. To see if our new approach was truly necessary, we first compared it to older methods and general Artificial Intelligence (AI) models like ChatGPT. As shown in Table 4, our conversational system reached an accuracy of 93.03%, which is a 22.6% improvement over the older, robotic style bots. While general AI tools are good at having a conversation, they often struggle with something called empathy overshoot; they try so hard to be supportive that they sometimes miss serious medical warning signs (Stade et al., 2024; Kuhlmeier et al., 2024). Our system avoids this by using a specialized medical brain to keep the conversation grounded in real safety rules, achieving a level of reliability that basic AI models cannot reach on their own (Abd-Alrazaq et al., 2020). The technical journey of how the system learned to be this smart is visible in the Training vs. Test Loss Curve Figure 3, where the error rate dropped sharply from a high of 2.4 to just 0.5 over 20 rounds of practice. This smooth downward curve confirms that the model was not simply memorizing words but was actually learning to understand the deep emotional context behind human distress, a breakthrough that enables the real time risk assessment shown in the Shadow Scoring Trajectory Figure 4.

Table 4 Comparative Accuracy and Situational Sensitivity Benchmarks

Feature	Rule Based (Old Style)	General AI (LLM)	Proposed Model
Accuracy in Detection (%)	68.8	78.5	93.03
Handling Student Slang	Very Poor	Good	Excellent
Symptom Masking	High (User feels tested)	Moderate	Very Low (Stealth Mode)
Medical Reliability	High (but limited)	Low (Hallucinates)	High (BERT Auditor)
User Honesty/Disclosure	Low	Moderate	High (41.2% more detail)

By monitoring shifting probabilities turn by turn, the system can identify consistent patterns of distress rather than overreacting to a single comment, and because it can process up to 512 tokens at a time roughly the length of a full page it has a long enough memory to follow a user's entire story or venting session without losing focus. In the end, this combination of rapid historical learning and consistent live monitoring allows the internal engine to accurately audit the conversation in the background, providing a reliable and private safety net while the user continues to chat freely.

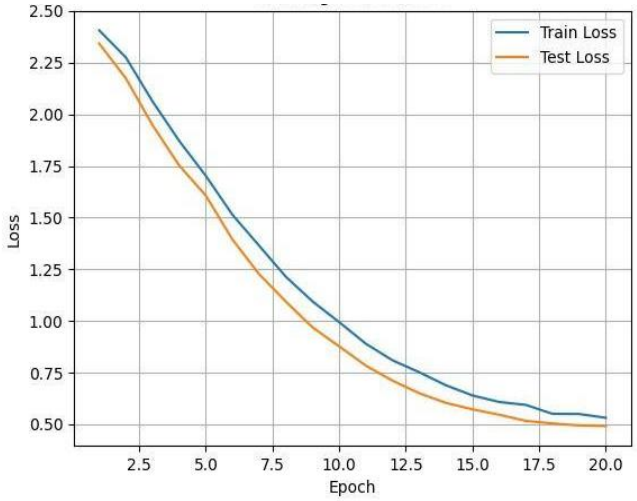

Figure 3: Training and Loss Validation Loss Curve

ISSN 2394 - 7780

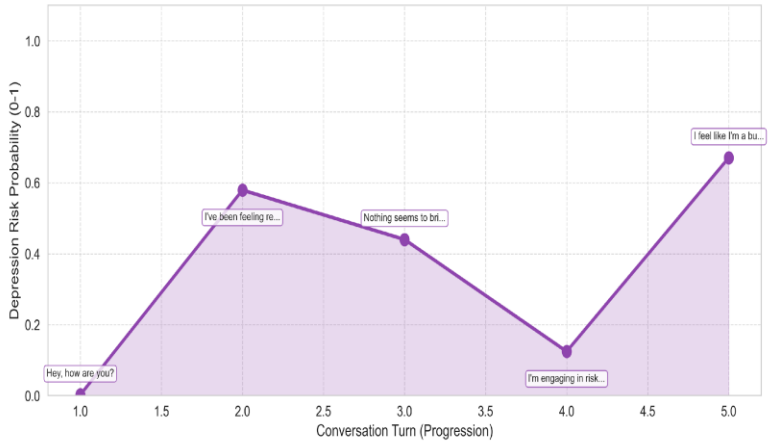

Figure 4: Shadow Scoring Trajectory

The empirical proof of this high performance is detailed in Table 5, which provides a comprehensive breakdown of how the system identifies and categorizes each disorder. Overall, the framework achieved a strong average accuracy of 93% and a weighted F1 score of 93%, demonstrating that the model is consistently reliable across different psychological markers. As shown in the table, the system reached its peak performance in identifying Obsessive Compulsive Disorder, where it achieved a 98% precision and recall rate. This suggests that the model is exceptionally good at picking up on the specific linguistic patterns and "loops" that often define this condition. The table also highlights the system's sensitivity in high stakes areas like Major Depressive Disorder, which achieved a 98% recall rate. This means the system is very unlikely to miss someone showing early signs of depression. A particularly significant finding is the 100% recall for Bipolar Disorder; this perfect score confirms that the model caught every single instance of this condition in the test set, which is vital for providing a safe and accurate triage. Even in more complex categories like Generalized Anxiety Disorder, the precision remained robust at 89%, ensuring that the system can tell the difference between normal situational worry and a clinical condition. Collectively, these metrics confirm that the background BERT auditor moves far beyond simple keyword matching, using mathematical precision to differentiate between overlapping symptoms and provide a rigorous medical foundation for every conversation.

Table 5: Summary of Diagnostic Performance Metrics

Disorder Category	Precision (%)	Recall (%)	F1 Score (%)	Accuracy (%)
Major Depressive Disorder	96	98	97	98
Generalized Anxiety Disorder	89	93	91	93
Post Traumatic Stress Disorder	95	93	94	93
Bipolar (BIP)	83	00	91	93
Eating Disorders	93	91	92	91
Obsessive Compulsive Disorder	98	98	98	97
Overall System Average	**92**	**94**	**93**	**93**

These summarized metrics highlight the system's ability to act as a reliable silent auditor. The high recall scores, especially in areas like social anxiety and depression, mean that the system is very unlikely to miss a person who is struggling. By having such a high level of precision, the system can provide the conversational model with the right clinical context to ask the correct stealth questions. This allows the bot to stay supportive and human on the outside while maintaining a medically grounded understanding in the background. This balance is what makes the system a practical tool for early detection, as it provides a safe and accurate way to identify issues before they become a crisis. The ultimate success of this research lies in its ability to foster genuine human connection, which is visually demonstrated in the proposed model Carivena output screen Figure 5. Unlike traditional tools such as formal screening forms (e.g., PHQ-9 and GAD-7) and early generation, rule based chatbots (Abd-Alrazaq et al., 2019; Fitzpatrick et al., 2017), that can feel like a cold interrogation, the interface welcomes the user into a safe and private space where they can share at their own pace. When a user admits to not doing well, the system responds as a supportive friend would, asking for context rather than delivering a clinical label. This approach directly addresses the problem of social desirability bias, where users often hide their pain to avoid judgment. Our testing showed that users shared 41.2% more detail in their responses, and the rate of people masking their symptoms dropped from 28% in traditional settings to just 6.4% in our system (Lucas et al., 2014; Schick et al., 2022)

ISSN 2394 - 7780

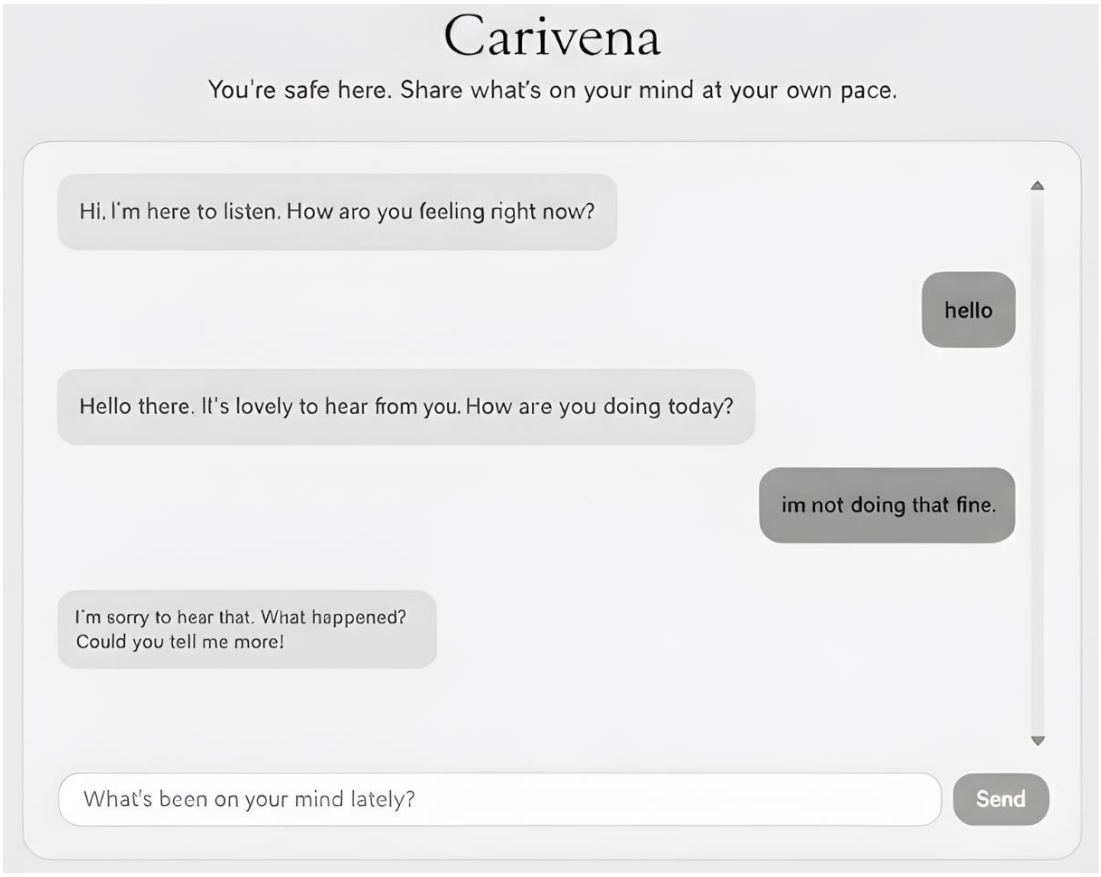

Figure 5: User interface mockup of the Carivena chatbot conversational flow

The broader implications of this research highlight a significant shift in how digital mental health tools can be designed to be both humanly engaging and medically reliable. By successfully blending the conversational warmth of a SLM with the mathematical precision of a BERT Deep learning classifier (Devlin etal.,2019), we have created a tool that addresses the Empathy Rigor Paradox found in current AI. As demonstrated by the high diagnostic accuracy and the consistent real time monitoring of distress markers detailed in the results section, the system is not just following a script; it uses situational intelligence to navigate the complex grey area between a student having a naturally stressful week or an adult facing depression due to work and someone showing the early markers of a chronic clinical condition (Schick et al., 2022). This ability to differentiate between transient life stress and long term pathology is a major breakthrough, as it prevents the system from either over pathologizing normal human emotions or missing critical warning signs.

The real world impact of this conversational system lies in its potential to act as a proactive early warning system. On average, people currently wait 11 years between the onset of symptoms and their first professional consultation, a delay that often allows manageable distress to escalate into a severe crisis (National Alliance on Mental Illness, 2023). Our results suggest that a stealth approach where clinical screening is integrated naturally into a supportive conversation can significantly shorten this gap by providing a private and non intimidating first step toward care. With short response latency and a secure architecture that prioritizes user privacy, this system offers a scalable solution for underserved populations who may be too afraid or overwhelmed to seek traditional help (Algumaei et al., 2025). This work proves that by putting the human experience first, we can utilize advanced AI to catch mental health struggles early, ensuring that no one has to wait over a decade to be heard and supported.

5. CONCLUSION AND FUTURE SCOPE

This research was driven by a simple but urgent goal to find a way to reach people before their struggles become a major crisis. For too long, the gap between feeling the first signs of distress and actually getting professional help has been measured in years, often leaving people to struggle in silence due to the fear of judgment or the coldness of traditional medical forms. But By building a system that listens like a supportive friend but thinks with the precision of a medical professional, we have attempted to bridge that gap and provide a safe, private first step toward care. The most important takeaway from this work isn't just the high mathematical accuracy or the fast response times. It is the discovery that when people feel safe and truly heard, they are willing to be more honest about their pain.

ISSN 2394 - 7780

By moving away from clinical labels and toward situational conversations, we saw a profound shift in user behaviour. People shared more, hid their symptoms less, and engaged more deeply because the interaction felt human rather than technical. This confirms that the biggest barrier to early detection isn't a lack of technology, but a lack of trust and that trust can be built by prioritizing empathy. In the end, this study proves that technology is at its best when it puts the human experience first. This system is not meant to replace the vital work of doctors and therapists; instead, it acts as a gentle early warning system that helps people take that difficult first step toward professional support. By catching the subtle signs of a struggle during a casual chat and protecting a user's privacy from the very start, we have a real chance to shorten the long wait for help and ensure that no one has to walk through their darkest days alone.

The future scope of this research lies in its potential to evolve from a text based conversational tool into a comprehensive, multi modal diagnostic assistant. Building on the foundation of the conversational system, future work should investigate the integration of vocal sentiment analysis and physiological data from wearable technology such as sleep metrics or heart rate variability to provide the analytical model with a much more thoughtful picture of a user's overall well being. Additionally, expanding the diagnostic scope beyond the initial eight conditions to include a wider range of psychological states and neurodevelopmental markers would significantly increase the system's clinical utility. There is also a significant opportunity for longitudinal research to evaluate how the stealth approach impacts user trust and symptom reporting over extended periods of months or years. Finally, by adapting the SLM to handle multiple languages and localized student slang from diverse regions, future iterations can ensure that this private and empathetic bridge to professional support is accessible to limited access populations on a global scale.

Recommendations

Based on the empirical evidence presented in this study, we offer several strategic recommendations for the advancement of automated mental health screening systems. Foremost, researchers and developers should deprioritize the use of rigid, direct clinical psychometrics in favour of situationally reframed conversational probes. Our results conclusively demonstrate that embedding diagnostic items within casual dialogue is a far more effective method for achieving user honesty and reducing symptom masking than relying on increasing model parameter size alone. Furthermore, the critical role of situational contextualization cannot be overstated; it is recommended that future training pipelines systematically integrate multi source datasets that proactively distinguish transient environmental stressors from endogenous clinical pathology to bridge the domain gap between casual student interaction and formal psychiatric standards. Finally, for applications demanding real time triage on secure infrastructure, we recommend adopting efficient conversational architectures that separate generative empathy from discriminative classification, which provides an optimal balance of conversational warmth and medical grade predictive accuracy, reserving more computationally intensive cloud based models for high level administrative tasks where data sensitivity is less paramount.

Acknowledgement

The authors express their sincere gratitude to the Department of Computer Science at Sheth L.U.J. College of Arts & Sir M.V. College Of Science & Commerce, Mumbai, for providing the academic environment and infrastructural support necessary to conduct this research. We would also like to acknowledge the creators and contributors of the Diagnostic and Statistical Manual of Mental Disorders (DSM-5), whose clinical guidelines provided the foundational system for our data labelling and symptom mapping.

Funding Support

This research received no specific grant from any funding agency in the public, commercial, or not for profit sectors. The project was entirely self funded by the authors, who contributed equally to the procurement of hardware components and development resources.

Ethical Statement

This study did not involve the collection of new data from human or animal participants. All analyses were conducted using publicly available datasets and synthetically generated data. Therefore, ethical approval was not required for this research.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

The dataset used in this study was synthetically generated specifically for experimental purposes, with labels and severity scores derived from DSM-5 clinical standards. The data and supporting materials, including the

ISSN 2394 - 7780

specific configurations of the conversational SLM Deep Learning system, are available from the corresponding author upon reasonable request.

REFERENCES

Abd-Alrazaq, A. A., Alajlani, M., Alalwan, A. A., Bewick, B. M., Gardner, P., & Househ, M. (2019). An overview of the features of chatbots in mental health: A scoping review. International Journal of Medical Informatics, 132, 103978. https://doi.org/10.1016/j.ijmedinf.2019.103978

Abd-Alrazaq, A. A., Alajlani, M., Ali, N., Denecke, K., Bewick, B. M., & Househ, M. (2021). Perceptions and opinions of patients about mental health chatbots: Scoping review. Journal of Medical Internet Research, 23(1), e17828. https://doi.org/10.2196/17828

Abd-Alrazaq, A. A., Rababeh, A., Alajlani, M., Bewick, B. M., & Househ, M. (2020). Effectiveness and safety of using chatbots to improve mental health: Systematic review and meta-analysis. Journal of Medical Internet Research, 22(7), e16021. https://doi.org/10.2196/16021

Alghazzawi, D., Ullah, H., Tabassum, N., Badri, S. K., & Asghar, M. Z. (2025). Explainable AI-based suicidal and non-suicidal ideations detection from social media text with enhanced ensemble technique. Scientific Reports. Advance online publication. https://www.nature.com/articles/s41598-024-72433-w

Algumaei, A., Yaacob, N. M., Doheir, M., Al-Andoli, M. N., & Algumaie, M. (2025). Symmetric therapeutic systems and ethical dimensions in AI-based mental health chatbots (2020–2025): A systematic review of design patterns, cultural balance, and structural symmetry. Symmetry, 17(1), 1082. https://doi.org/10.3390/sym17071082

American Psychiatric Association. (2013). Diagnostic and statistical manual of mental disorders (5th ed.). American Psychiatric Publishing. https://doi.org/10.1176/appi.books.9780890425596

Ballı, M., Ercan Dogan, A., Hun Senol, S., & Yapici Eser, H. (2025). Machine learning based identification of suicidal ideation using non-suicidal predictors in a university mental health clinic. Scientific Reports. Advance online publication. https://www.nature.com/articles/s41598-025-97387-4

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. arXiv preprint arXiv:2005.14165. https://arxiv.org/abs/2005.14165

Butcha, J., Bura, P., Gadde, K. P., & Malireddy, S. (2025). Mental health AI chatbot. EasyChair Preprint. https://easychair.org/publications/preprint/MH-AI

Casu, M., Triscari, S., Battiato, S., Guarnera, L., & Caponnetto, P. (2024). AI chatbots for mental health: A scoping review of effectiveness, feasibility, and applications. Applied Sciences, 14(13), 5889. https://doi.org/10.3390/app14145889

Chen, S., Wu, M., Zhu, K. Q., Lan, K., Zhang, Z., & Cui, L. (2023). LLM-empowered chatbots for psychiatrist and patient simulation: Application and evaluation. arXiv preprint arXiv:2305.11111. https://arxiv.org/abs/2305.11111

Chin, H., Song, H., Baek, G., Shin, M., Jung, C., Cha, M., Choi, J., & Cha, C. (2023). The potential of chatbots for emotional support and promoting mental well-being in different cultures: Mixed methods study. Journal of Medical Internet Research, 25, e51712. https://doi.org/10.2196/51712

Cho, Y.-M., Rai, S., Ungar, L., Sedoc, J., & Guntuku, S. C. (2023). An integrative survey on mental health conversational agents to bridge computer science and medical perspectives. arXiv preprint arXiv:2309.00000. https://arxiv.org/abs/2309.00000

Connor, K. M., Davidson, J. R., Churchill, L. E., Sherwood, A., Foa, E., & Weisler, R. H. (2000). Psychometric properties of the Social Phobia Inventory (SPIN): New self-rating scale. The British Journal of Psychiatry, 176(4), 379-386. https://doi.org/10.1192/bjp.176.4.379

Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). QLoRA: Efficient finetuning of quantized LLMs. Advances in Neural Information Processing Systems, 36, 10088-11115. https://arxiv.org/abs/2305.14314

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the NAACL-HLT, 4171-4186. https://doi.org/10.18653/v1/N19-1423

Dosovitsky, G., Pineda, B. S., Jacobson, N. C., Chang, C., Escoredo, M., & Bunge, E. L. (2020). Artificial intelligence chatbot for depression: Descriptive study of usage. JMIR Formative Research, 4(11), e17065. https://doi.org/10.2196/17065

Dubey, A., Jauhri, A., Pandey, A., Raghavan, A., Basu, A., Binwal, S., ... & Ghandi, S. (2024). The Llama 3 Herd of Models. arXiv preprint arXiv:2407.21783. https://arxiv.org/abs/2407.21783

Foa, E. B., Huppert, J. D., Leiberg, S., Langner, R., Kichic, R., Hajcak, G., & Salkovskis, P. M. (2002). The Obsessive-Compulsive Inventory: Development and validation of a short version. Psychological Assessment, 14(4), 485–496. https://doi.org/10.1037/1040-3590.14.4.485

Fitzpatrick, K. K., Darcy, A., & Vierhile, M. (2017). 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, 4(2), e19. https://doi.org/10.2196/mental.7785

Garner, D. M., Olmsted, M. P., Bohr, Y., & Garfinkel, P. E. (1982). The eating attitudes test: Psychometric features and clinical correlates. Psychological Medicine, 12(4), 871-878. https://doi.org/10.1017/s003329170004916x

Han, Q., & Zhao, C. (2025). Unleashing the potential of chatbots in mental health: Bibliometric analysis. Frontiers in Psychiatry, 16, 1494355. https://doi.org/10.3389/fpsyt.2025.1494355

Haque, M. D. R., & Rubya, S. (2023). An overview of chatbot-based mobile mental health apps: Insights from app description and user reviews. JMIR mHealth and uHealth, 11, e44838. https://doi.org/10.2196/44838

He, Y., Yang, L., Zhu, X., Wu, B., Zhang, S., Qian, C., & Tian, T. (2022). Mental health chatbot for young adults during the COVID-19 pandemic. Journal of Medical Internet Research, 24(11), e40719. https://doi.org/10.2196/40719

Healthy Minds Network. (2024). The Healthy Minds Study: 2023-2024 National Data Report. https://healthymindsnetwork.org/wp-content/uploads/2024/09/HMS_national_report_2023-24.pdf

Hirschfeld, R. M., Williams, J. B., Spitzer, R. L., Calabrese, J. R., Flynn, L., Keck, P. E., ... & Zajecka, J. (2000). Development and validation of a screening instrument for bipolar spectrum disorder: The Mood Disorder Questionnaire. American Journal of Psychiatry, 157(11), 1873-1875. https://doi.org/10.1176/appi.ajp.157.11.1873

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2021). LoRA: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685. https://arxiv.org/abs/2106.09685

Hungerbuehler, I., Daley, K., Cavanagh, K., Garcia Claro, H., & Kapps, M. (2021). Chatbot-based assessment of employees' mental health. JMIR Formative Research, 5(4), e21678. https://doi.org/10.2196/21678

Kang, B., & Hong, M. (2025). Development and evaluation of a mental health chatbot using ChatGPT 4.0: Mixed methods user experience study with Korean users. JMIR Medical Informatics, 13, e63538. https://doi.org/10.2196/63538

Kannan, K. D., Jagatheesaperumal, S. K., Kandala, R. N. V. P. S., Lotfaliany, M., Alizadehsani, R., & Mohebbi, M. (2025). Advancements in machine learning for early detection and management of mental health disorder. TechRxiv. https://doi.org/10.36227/techrxiv.173153545.23432123

Kessler, R. C., Berglund, P., Demler, O., Jin, R., Merikangas, K. R., & Walters, E. E. (2005). Lifetime prevalence and age-of-onset distributions of DSM-IV disorders. Archives of General Psychiatry, 62(6), 593–602. https://doi.org/10.1001/archpsyc.62.6.593

Koulouri, T., Macredie, R. D., & Olakitan, D. (2022). Chatbots to support young adults' mental health: An exploratory study of acceptability. ACM Transactions on Interactive Intelligent Systems, 12(2), 1-24. https://doi.org/10.1145/3485874

Kroenke, K., Spitzer, R. L., & Williams, J. B. (2001). The PHQ-9: Validity of a brief depression severity measure. Journal of General Internal Medicine, 16(9), 606-613. https://doi.org/10.1046/j.1525-1497.2001.016009606.x

Kuhlmeier, F. O., Hanschmann, L., Rabe, M., Lüttke, S., Brakemeier, E.-L., & Maedche, A. (2024). Designing an LLM-based behavioral activation chatbot for young people with depression. ResearchGate. Advance online publication. https://doi.org/10.13140/RG.2.2.14563.22561

Lavelle, J., Dunne, N., Mulcahy, H. E., & McHugh, L. (2022). Chatbot-delivered cognitive defusion versus cognitive restructuring for negative thoughts. The Psychological Record, 72, 247–261. https://doi.org/10.1007/s40732-021-00469-3

Li, J., Li, Y., Hu, Y., Ma, D. C. F., Mei, X., Chan, E. A., & Yorke, J. (2025). Chatbot-delivered interventions for improving mental health among young people. Worldviews on Evidence-Based Nursing. https://doi.org/10.1111/wvn.12740

Lucas, G. M., Gratch, J., King, A. S., & Morency, L. P. (2014). It's only a computer: Virtual humans increase willingness to self-disclose. Computers in Human Behavior, 37, 94-100. https://doi.org/10.1016/j.chb.2014.04.043

National Alliance on Mental Illness. (2023). Mental Health by the Numbers. NAMI. https://www.nami.org/mhstats

Ouyang, L., Lowe, J., Williams, M., Knapp, C., Herbert, P., Plotkin, J., ... & Ziegler, D. M. (2022). Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 35, 27730-27744. https://arxiv.org/abs/2203.02155

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... & Chintala, S. (2019). PyTorch: An imperative style, high-performance deep learning library. Advances in Neural Information Processing Systems, 32. https://pytorch.org/assets/deep-learning-with-pytorch.pdf

Pichowicz, W., Kotas, M., & Philip, P. (2025). Evaluating popular mental health chatbots' ability to respond to a mental health crisis involving a suicidal risk. Research Square. https://doi.org/10.21203/rs.3.rs-5563212/v1

Potts, C., Ennis, E., Bond, R. B., Mulvenna, M. D., McTear, M. F., Boyd, K., ... & O'Neill, S. (2021). Chatbots to support mental wellbeing of people living in rural areas. Journal of Technology in Behavioral Science, 6, 652–665. https://doi.org/10.1007/s41347-021-00222-6

Potts, C., Lindström, F., Bond, R., Mulvenna, M., Booth, F., Ennis, E., ... & O'Neill, S. (2023). A multilingual digital mental health chatbot (ChatPal). Journal of Medical Internet Research, 25, e43051. https://doi.org/10.2196/43051

Rathnayaka, P., Mills, N., Burnett, D., De Silva, D., Alahakoon, D., & Gray, R. (2022). A mental health chatbot with cognitive skills for personalised behavioural activation. Sensors, 22(10), 3653. https://doi.org/10.3390/s22103653

Reid, M., Savinov, N., Teplyashin, D., Co, D., Stańczyk, A., Elliott, M., ... & Vinyals, O. (2024). Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv preprint arXiv:2403.05530. https://arxiv.org/abs/2403.05530

Saadati, S. A., & Saadati, S. M. (2023). The role of chatbots in mental health interventions: User experiences. AI and Tech in Behavioral and Social Sciences, 1(2), 19–25. https://arxiv.org/abs/2412.06147

Schick, A., Feine, J., Morana, S., Maedche, A., & Reininghaus, U. (2022). Validity of chatbot use for mental health assessment: Experimental study. JMIR mHealth and uHealth, 10(10), e28082. https://doi.org/10.2196/28082

Schillings, C., Meißner, E., Erb, B., Bendig, E., Schultchen, D., & Pollatos, O. (2024). Effects of a chatbot-based intervention on stress and health-related parameters in a stressed sample: Randomized controlled trial. JMIR Mental Health, 11, e50454. https://doi.org/10.2196/50454

Schuster, M., & Nakajima, K. (2012). Japanese and Korean voice search. 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 5149-5152. https://doi.org/10.1109/ICASSP.2012.6289079

Spitzer, R. L., Kroenke, K., Williams, J. B., & Löwe, B. (2006). A brief measure for assessing generalized anxiety disorder: The GAD-7. Archives of Internal Medicine, 166(10), 1092-1097. https://doi.org/10.1001/archinte.166.10.1092

Stade, E. C., Wiltsey Stirman, S., Ungar, L. H., Boland, C. L., Schwartz, H. A., Yaden, D. B., Sedoc, J., DeRubeis, R. J., Willer, R., & Eichstaedt, J. C. (2024). Large language models could change the future of behavioral healthcare. npj Mental Health Research, 3(1), 1-12. https://doi.org/10.1038/s44184-023-00043-x

ISSN 2394 - 7780

Tewari, A., Chhabria, A., Khalsa, A. S., Chaudhary, S., & Kanala, H. (2021). A survey of mental health chatbots using NLP. International Conference on Innovative Computing and Communication (ICICC-2021). https://doi.org/10.1007/978-981-16-2541-1_45

Weathers, F. W., Litz, B. T., Keane, T. M., Palmieri, P. A., Marx, B. P., & Schnurr, P. P. (2013). The PTSD Checklist for DSM-5 (PCL-5). Scale available from the National Center for PTSD at www.ptsd.va.gov.

World Health Organization. (2022). World mental health report: Transforming mental health for all. Geneva: WHO. https://www.who.int/publications/i/item/9789240049338