ISBN: 978-628-96613-3-0. ISSN: 2414-6390. Digital Object Identifier: https://dx.doi.org/10.18687/LACCEI2026.1.1.241

Model of a GPT-4 and Natural Language Processing Driven Chatbot for Mental Health Strategies Among University Students in Peru

Jesus Luna[1]; Nemberk Jara[2]; Keo Pozo[3]
[1,2,3]*Universidad Peruana de Ciencias Aplicadas, Perú, pcapjlun@upc.edu.pe, u202120739@upc.edu.pe, u202120644@upc.edu.pe*

Abstract*– Youth depression and suicide represent an increasingly critical public health challenge in Peru, with a significant impact on the university population. This paper presents a technological model based on an emotional-support chatbot that integrates the GPT-4 architecture and natural language processing (NLP) techniques. The approach combines sentiment analysis with the generation of empathetic responses, complemented by the delivery of educational resources. A quasi-experimental pretest–posttest design without a control group was conducted with a sample of 30 university students over a one-month intervention period. The evaluation included changes in depressive symptoms using the PHQ-9 scale, usability through the System Usability Scale (SUS), and perceived empathy and response performance. The results show an observed reduction in PHQ-9 scores, high usability, and positive user perception of empathy, suggesting that the chatbot provides a safe and accessible space for emotional support. However, these findings should be interpreted cautiously due to the limited sample size and absence of a control group. This proposal aims to strengthen preventive mental health strategies in the Peruvian university context through the use of innovative digital tools.

***Keywords*-- mental health, youth, chatbot, natural language processing, GPT-4**

I. INTRODUCTION

Youth mental health has become one of the most pressing challenges in global public health. The World Health Organization (WHO) estimates that approximately 15% of young people experience some type of mental disorder before the age of 19, resulting in significant impacts on their academic, social, and emotional development [1]. These conditions are closely linked to suicide, which has emerged as one of the leading causes of death among individuals aged 15 to 29 [2]. This trend is also evident in Peru, where reports from the Ministry of Health (MINSA) indicate a sustained increase in suicide attempts over the past decade, particularly among adolescents and young adults [3].

The emergence of large language models (LLMs), particularly GPT-4 (Generative Pre-trained Transformer 4), has opened new possibilities for developing chatbots that are more empathetic, adaptive, and contextually sensitive. Unlike traditional systems, GPT-4 enables coherent and natural interactions, with the ability to recognize emotional cues and tailor responses to the user's situation. Recent studies show that GPT-4-based chatbots can deliver more fluid and personalized conversations while incorporating clinical screening tools and risk-detection algorithms [4].

This paper proposes and evaluates a mental-health chatbot model for Peruvian university students, grounded in the integration of GPT-4 with Natural Language Processing (NLP) techniques. The model seeks to address the limitations of previous systems by implementing adaptive and culturally sensitive conversational flows, integrating standardized clinical instruments such as the Patient Health Questionnaire-9 (PHQ-9) for depressive-symptom screening, and enabling referral to professional support services. A pilot study using a quasi-experimental pretest–posttest design was conducted to assess feasibility, usability, and preliminary impact.

II. RELATED WORK

Kim et al. [4] developed and evaluated HoMemeTown Dr. CareSam, a GPT-4–based bilingual mental-health chatbot featuring gratitude journaling, emotion detection, and suicide-risk management. In a pilot study with Korean youth, the system received high ratings for support (9.0/10) and empathy (8.7/10), outperforming both other language models and existing therapeutic chatbots ($F = 12.94$; $p < .001$). Key limitations included limited personalization and response latency. The authors concluded that the approach is promising as an accessible mental-health support tool, although broader clinical trials and linguistic refinements are needed.

Karkosz et al. [5] evaluated Fido, a Polish therapeutic chatbot based on cognitive-behavioral therapy, in a trial involving young adults with subclinical symptoms of depression and anxiety. After two weeks, users showed significant reductions in depression (Patient Health Questionnaire-9 (PHQ-9)decreased from 9.3 to 6.8; $F = 34.18$; $p < .001$) and anxiety (State–Trait Anxiety Inventory (STAI) decreased from 42.1 to 37.5; $F = 25.87$; $p < .001$), as well as improvements in well-being (Satisfaction With Life Scale (SWLS) increased from 18.2 to 20.4; $F = 13.59$; $p < .001$). No differences were observed compared with the self-help control group, although frequent users reported decreased loneliness (Revised UCLA Loneliness Scale (R-UCLA) from 45.0 to 42.3; $F = 7.41$; $p = .01$). Limitations included intention-recognition errors and the strong effectiveness of the control materials. Overall, the results demonstrate the chatbot's potential as an accessible support resource.

Ulrich et al. [6] developed MISHA, a chatbot offering psychoeducation, relaxation sessions, and mindfulness for Swiss university students. The intervention produced significant reductions in stress, depression, and psychosomatic symptoms (d = −0.36 to −0.60), although no improvements were observed in anxiety or active coping. Limited personalization and insufficient testing among larger populations constrained its applicability.

Potts et al. [7] evaluated ChatPal, a multilingual chatbot deployed in rural regions across five European countries, integrating gratitude, mindfulness, and mood tracking. The study reported slight increases in well-being, particularly among younger and more frequent users; however, the effects were not statistically significant ($p > .42$). Technical issues and limited personalization were major constraints.

Rathnayaka et al. [8] introduced an AI-supported behavioral activation chatbot designed for personalized interventions and remote monitoring. A small-scale pilot indicated that personalization enhanced continuity of emotional support and increased perceived effectiveness. Nonetheless, the study was limited by a small sample and the absence of robust clinical comparisons.

Eltahawy et al. [9] compared Woebot, ELIZA, journaling, and psychoeducation among 65 young participants, finding that Woebot did not outperform the other technologies, while ELIZA and journaling showed greater symptom reductions. The study was restricted by a small sample and findings contrary to prior evidence, underscoring the need for greater methodological rigor before equating chatbots with psychotherapy.

Kang and Hong [10] tested Woebot and Happify among 27 Korean university students after the pandemic. Although reductions in depression and loneliness were observed, they did not reach statistical significance ($p > .35$). High attrition and the small sample size limited the validity of the results.

Kaywan et al. [11] developed DEPRA, a chatbot built on Dialogflow that administered clinical scales (SIGH-D, IDS-C) for early depression screening. In a trial with 50 participants, the chatbot achieved classification performance comparable to traditional instruments and obtained a 79% user satisfaction rate.

Maciejewski and Smoktunowicz [12] designed Stressbot, a Messenger-based chatbot aimed at promoting self-efficacy and reducing stress in students. Results showed significant improvements in stress ($d = -0.33$) and self-efficacy ($d = 0.50$) after the intervention, although the effects dissipated after one month. Limited personalization and sustainability reduced long-term impact. The use of popular platforms demonstrates feasibility, provided redesigns are implemented to maintain benefits.

He et al. [13] developed XiaoE, a Cognitive Behavioral Therapy (CBT)-based chatbot deployed with Chinese university students during the pandemic, achieving significant reductions in depressive symptoms ($d = 0.51$ to 0.31; $p < .001$) and improvements in therapeutic alliance and acceptability. The short one-week intervention and lack of long-term follow-up limited the findings. The study supports the feasibility of CBT chatbots for youth but highlights the need for replicated and longer trials.

Goyal et al. [14] proposed MindLift, a multimodal assessment platform combining text, voice, and facial-expression analysis using Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and NLP for real-time student monitoring. Initial findings showed strong student willingness to seek help and improvements in emotional regulation, supported by solid accuracy, recall, and F1-score metrics. Ethical concerns and algorithmic bias remain challenges. The initiative appears promising and scalable for educational institutions.

Obadinma et al. [15] developed FAIIR, a transformer-based assistant trained on 780,000 crisis-related youth conversations, achieving an area under the curve (AUC) of 94% and 81% recall, with 90.9% agreement between expert assessments and model predictions. Lower precision in rare categories and the need for model refinement were noted as limitations. The findings demonstrate strong potential for supporting triage and prioritization of urgent youth mental-health cases.

Kim et al. [16] analyzed *Luda Lee*, a social chatbot used by Korean university students over a four-week period, reporting significant reductions in loneliness ($p = .02$) and social anxiety ($p = .01$). Factors such as resilience and self-disclosure toward the chatbot predicted better outcomes. However, inconsistent responses and excessive enthusiasm affected user experience. The study illustrates how social chatbots can complement mental-health interventions by providing empathetic companionship.

Chin et al. [17] examined 152,783 interactions with *SimSimi* to analyze cultural differences in emotional-chatbot use across Eastern and Western users. Eastern users expressed more intense emotions ($p = .01$), whereas Western users more frequently discussed sensitive topics such as mental health or death ($p < .001$). The analysis was limited to textual content and lacked clinical measures. The findings highlight chatbots as culturally differentiated safe spaces for expressing emotional vulnerability.

Chong et al. [18] implemented a digital suicide-risk notification system that sent immediate alerts to clinicians for individuals aged 16–25 at critical risk. The system resolved 76% of alerts within an average of 1.9 days, with 60% resulting in completed safety plans. Limitations included low personalization and delays in some responses. The solution demonstrated the capacity to accelerate care and prevent symptom escalation in overloaded clinical services.

Yeo et al. [19] evaluated *Acceset*, a peer-support digital platform for university students, as an alternative for addressing high rates of depression and anxiety in young adults. Significant improvements were found in self-compassion, mindfulness, and depressive symptoms ($t = 3.44$; $p < .001$). Reliance on peer training and limited generalizability were key constraints. Peer-based digital support appears to be a sustainable approach to psychological well-being.

Meyerhoff et al. [20] developed an automated SMS-based safety-planning system co-designed with youth who had experienced suicidal ideation. Users emphasized the importance of transparency, privacy, and message personalization, as well as integrating support through social networks. The study lacked field-based empirical validation, limiting its scope. The findings underscore the value of youth co-design in developing safe and trustworthy digital tools.

Niu et al. [21] applied random-forest models to 4,070 youth survivors of childhood sexual abuse to differentiate suicide-related phases: ideation, planning, and attempt. Depression and non-suicidal self-injury emerged as common predictors, while self-compassion was a key factor distinguishing attempts. The model showed low accuracy in separating attempts from plans. Identifying phase-specific predictors provides a path toward more targeted interventions.

Cliffe et al. [22] conducted a workshop with experts, youth, and technical staff to improve *Tellmi*, a peer-support digital platform. Recommendations included clearer objectives, personalized support, and strategies to protect peer responders. The study was limited by a small and unrepresentative sample. Interdisciplinary and user collaboration proved essential for creating safer tools for high-risk youth.

Rainbow et al. [23] evaluated *Beyond Now* in 610 users with a history of suicidal ideation, measuring changes over three months. Significant reductions in ideation and improvements in coping were observed ($p < .001$), although usage time and co-creation factors did not affect outcomes. Heavy reliance on self-reports and lack of data on suicide attempts were limitations. Personalized digital safety plans strengthened coping perceptions and reduced ideation over time.

Tong et al. [24] assessed a self-care–themed chatbot among 285 participants from the general population, focusing on emotional regulation and stress management. Results showed improvements in self-care intentions, mental-health literacy, and depressive symptoms (Cohen's $d = -0.26$; $p = .004$), though effects diminished after one month. Limited personalization and sustainability constrained long-term impact. The study confirms the preventive potential of chatbots while emphasizing the need for greater interactivity and generative intelligence.

Weiner et al. [25] evaluated *Vira* in 73 youth with depressive symptoms and obesity risk, comparing an unguided version with a coaching-supported version. Both modalities reduced depression and anxiety, with greater improvements in the coaching group ($d = 0.45$ and $d = 0.50$). The small sample size and lack of an active control group limited generalizability. Combining digital monitoring with human support showed enhanced intervention effects.

Torok et al. [26] evaluated *LifeBuoy*, a Dialectical Behavior Therapy (DBT)-based mobile application for 455 youth who had experienced suicidal ideation. The app significantly reduced ideation levels ($d = 0.45$ post-test; $d = 0.34$ at three months), though no changes were observed in secondary symptoms. High attrition (40%) and insufficient statistical power for secondary outcomes were key limitations. The study demonstrates that suicide-specific digital interventions can be effective but require adherence strategies and expanded clinical validation.

Litvin et al. [27] analyzed *eQuoo*, a gamified psychological-skills application used by 1,165 students across 180 universities. Significant improvements were observed in resilience and reductions in anxiety and depression ($p < .001$), with a 42% higher retention rate compared to controls. Lack of long-term follow-up and difficulty isolating the gamification effect were limitations. Gamification appears promising for enhancing adherence and effectiveness in digital interventions for students.

Wrightson-Hester et al. [28] developed *MYLO*, a mental-health chatbot based on the Method of Levels (MoL), co-designed with youth aged 16–24 and tested in a two-week pilot. The system achieved large reductions in problem-related distress ($d = 1.50$) and moderate improvements in conflict reorganization, though repetitive questions and limited language comprehension reduced usability. The authors conclude that MYLO is feasible and well-received, with potential to support youth mental health pending further clinical trials and personalization enhancements.

Park and Kim [29] analyzed factors influencing the adoption of *MyMentalPocket*, a digital mental-health platform featuring the emotional-support chatbot "Pocky," among 278 Korean students and university staff. Social interaction with the chatbot was the strongest predictor of usage intention ($\beta = 0.46$; $p < .001$), followed by perceived usefulness ($\beta = 0.37$; $p < .001$), while ease of use was not significant. Women, younger participants, and those with higher depressive symptoms tended to engage more. The findings highlight the importance of perceived empathy and companionship in chatbot acceptance for depression management.

Koulouri et al. [30] explored the acceptability of mental-health chatbots among young adults through a university survey, literature review, and counselor interviews. In the survey ($n = 150$), 53% reported mental-health issues in the past year, with high levels of stress (80%) and depression (50%), yet nearly half did not know where to seek support. Only 37% felt comfortable discussing mental health with a chatbot, although most had used assistants like Siri or Alexa, indicating technological familiarity. Counselors viewed chatbots as useful for initial contact and complementary support, provided they integrate interactivity, personalization, and referral to professional services. The study concludes that chatbots hold acceptance potential in this age group, though concerns about privacy, trust, and clinical effectiveness persist.

III. ARCHITECTURE MODEL

A. Theoretical Foundations of Conversational Architectures

Artificial Intelligence (AI)-driven conversational chatbots represent a significant evolution in digital mental-health interventions. These systems integrate large-scale language models (LLMs) and Natural Language Processing (NLP) techniques to generate empathetic, personalized, and culturally sensitive interactions. Kang and Hong [4] note that the development of GPT-4–powered chatbots demonstrates remarkable potential for delivering accessible emotional support by combining advanced linguistic capabilities with an architecture designed to prioritize user privacy and psychological risk detection.

B. Design and Configuration of the Proposed Architecture

Fig. 1 presents the proposed physical architecture, inspired by the modular approaches outlined by Kang and Hong [4] and Rathnayaka et al. [8], who emphasize the importance of distributed infrastructures that integrate intelligent processing, secure storage, and an accessible interaction layer. In this work, the architecture is adapted to the Peruvian university context through a lightweight and scalable configuration consisting of a mobile application developed in Flutter, a backend implemented in FastAPI using Python, a PostgreSQL database composed of nine entities, the GPT-4 conversational inference service hosted on the OpenAI platform, and the Google Natural Language Application Programming Interface (API). The entire solution is deployed on Google Cloud Platform (GCP).

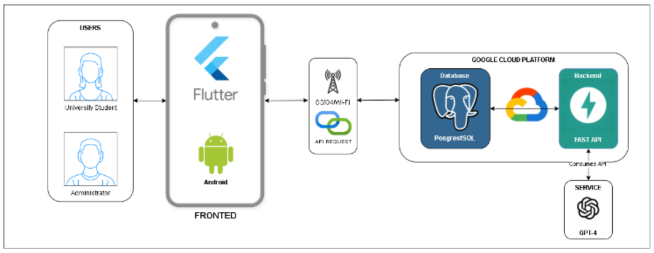

Fig. 1. Proposed physical architecture of the chatbot

C. Application Flow

Fig. 2 illustrates the overall workflow of the chatbot, named SeaBot, describing the interactions between the user, the mobile application, and the intelligent backend through five main phases: (1) secure registration and authentication, (2) conversational interaction with GPT-4 for emotional support, (3) biweekly assessment using the PHQ-9 questionnaire, (4) rapid emotional logging via emojis, and (5) generation of emotional-evolution reports. Together, the system integrates Flutter, FastAPI, and GPT-4 within a secure and scalable cloud environment (GCP), providing a personalized experience for emotional support and psychological-well-being monitoring.

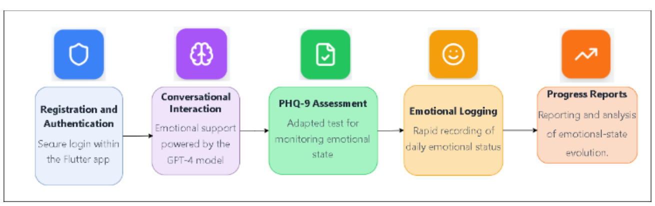

Fig. 2. System operating flow

IV. METHOD

A. Dataset

Table I presents the structure of the dataset used in this study. The dataset comprises approximately 2,000 user–chatbot interaction messages collected during the intervention period. The selection of this data source aims to build a representative corpus that reflects the linguistic, emotional, and behavioral patterns of Peruvian university students. In this study, the label "Negative" in the Sentiment Analysis field refers to user messages expressing emotions such as sadness, hopelessness, exhaustion, anxiety, or self-devaluation.

Table I
STRUCTURE OF THE DATASET USED FOR ANALYZING INTERACTIONS BETWEEN USERS AND THE SEABOT CHATBOT

Record Category	Description	Example Value
Identifier	Anonymous code	U-012
User input	Message sent to the chatbot	"I feel very exhausted and unmotivated."
Chatbot response	GPT-4-generated empathetic message	"I understand how you feel. Would you like to talk about what has been worrying you?"
Sentiment analysis	Emotional classification of user input	Negative
Assessment	PHQ-9 score	11

Dataset construction comprises two main components: conversational logs generated during chatbot usage and standardized psychological assessment results, such as the Patient Health Questionnaire-9 (PHQ-9). These data enable the analysis of the relationship between emotional interaction and the progression of depressive symptoms. Each record includes key variables to support processing and analysis: anonymous user identifier, user message, chatbot-generated response, detected sentiment label, and PHQ-9 score. This structure facilitates the application of performance, empathy, and safety metrics for evaluating model responses. To assess the reliability of sentiment classification, a subset of messages was manually annotated and compared with the automated labels generated by the Google NLP API. Standard evaluation metrics such as accuracy, precision, recall, and F1-score were computed.

B. Experimental Procedure and Study Design

The validation process for the SeaBot chatbot was conducted at the Peruvian University of Applied Sciences. The sample consisted of 30 students aged 18–25 presenting mild or moderate depressive symptoms. Participants interacted with the chatbot over a one-month period. The study aims to evaluate feasibility, usability, and preliminary psychological impact within the Peruvian cultural and educational context. A quasi-experimental single-group pretest–posttest design without a control group was implemented, using clinical instruments administered before and after the intervention. Cases identified with high risk (PHQ-9 $\geq$ 20) were immediately referred to face-to-face professional care. Throughout the study, all conversations were ethically monitored and anonymized exclusively for research purposes.

C. Evaluation and Validation Methods

The chatbot was evaluated across two primary dimensions: psychological effectiveness and technical performance. The objective was to determine whether chatbot use contributed to reductions in depressive symptoms and improvements in emotional well-being among university students.

1. Psychological Evaluation

Table II presents the Patient Health Questionnaire-9 (PHQ-9), adapted from Kroenke et al. [31]. The questionnaire was administered before and after the intervention period. The score difference (ΔPHQ-9) was analyzed using paired t-tests to compare pretest and posttest scores within the same group of participants. Descriptive statistics, including mean and standard deviation, were calculated. Statistical significance was evaluated using p-values, and effect size was estimated using Cohen's d.

Table II.
PHQ-9 scale and interpretation of depression severity

PHQ-9 Score	Severity Level	Clinical Interpretation
0–4	None–minimal	No significant depression
5–9	Mild	Initial symptoms; monitoring recommended
10–14	Moderate	Low to moderate clinical risk
15–19	Moderately severe	High risk
20–27	Severe	High risk of suicidal ideation

2. Technical and Usability Evaluation

Table III summarizes the technical and usability metrics used to assess semantic coherence, perceived empathy, response time, and overall system usability.

Table III
TECHNICAL, PSYCHOLOGICAL, AND USABILITY METRICS FOR THE SEABOT CHATBOT

Dimension	Metric	Description	Expected Threshold
Psychological effectiveness	ΔPHQ-9 (pre–post)	Difference between initial and final PHQ-9 scores	≥ 3-point reduction (clinical improvement)
	Percentage of users with reduction	Percentage of students who reduce at least 30% of their initial score.	≥30%
Technical performance	Semantic coherence	Fluency and contextual alignment of chatbot responses	≥ 80%.
	Perceived empathy	User-rated empathy of chatbot responses	≥ 80%.
	Average response time	Mean response-generation time (seconds)	≤ 3s.
Usability & satisfaction	System Usability Scale (SUS)	Standardized 10-item usability score	≥ 80%.
	Retention rate	(Active users at end ÷ total initial users) × 100	≥ 70%.

Descriptive statistics were used to summarize system performance and user perception metrics, including mean values and percentage distributions.

3. Sentiment Analysis Validation

To evaluate the performance of the sentiment analysis component, a subset of user messages was manually annotated and compared with the labels generated by the Google NLP API. Classification performance was assessed using standard metrics, including accuracy, precision, recall, and F1-score, based on the comparison between predicted and manually assigned labels.

V. ETHICAL CONSIDERATIONS AND DATA SECURITY

Given the sensitive nature of mental health data, the SeaBot system was designed with strict ethical and security measures. All user interactions were anonymized to protect personal identity, and no personally identifiable information was exposed during data processing.

Communication between the mobile application and backend services was secured using encrypted protocols (HTTPS with TLS), ensuring data confidentiality and integrity during transmission. User data was stored in a secure cloud-based environment (Google Cloud SQL), with controlled access and managed infrastructure protection.

From an ethical perspective, SeaBot does not provide medical diagnoses nor replace professional psychological care. Instead, it offers emotional support, psychoeducation, and guidance within a preventive framework. In cases where users exhibited high-risk indicators (e.g., PHQ-9 ≥ 20), the system provided recommendations to seek professional mental health services, supporting appropriate risk escalation.

To enhance safety in AI-generated responses, the GPT-4 model was guided through predefined conversational constraints and curated response patterns aligned with best practices in psychological support. These safeguards aimed to reduce the risk of inappropriate or harmful outputs during user interaction. All participants provided informed consent prior to participation, ensuring compliance with ethical standards for research involving human subjects.

VI. RESULTS

A. User Interaction and Engagement Metrics

Fig. 3 presents the general interaction metrics collected during the pilot phase, including the total number of sessions, average conversation duration, and average messages per session. During the first week, participants engaged in an average of 2.1 sessions, with a mean duration of 8 minutes and 12.7 messages per session. In the second week, a slight decrease was observed, with an average of 2.0 sessions per user and a reduction in conversation duration to 6.9 minutes, while the number of messages per session remained relatively stable at 16.1. In the third week, an average of 1.9 sessions per user was recorded, with an average duration of 5.6 minutes per session and 10.4 messages per session. Finally, in the last week, an average of 1.5 sessions, 4.5 minutes per session, and 8.6 messages per session were observed. These trends suggest consistent user participation throughout the intervention period.

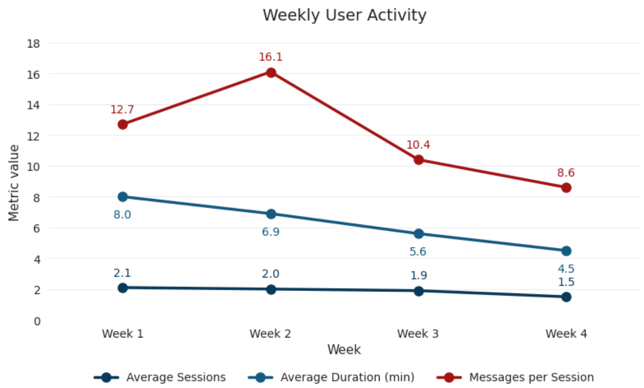

Fig. 3. Weekly user activity metrics, including average sessions, duration, and messages per session.

B. *Emotional Detection and Sentiment Analysis*

Fig. 4. illustrates the distribution of emotional categories identified through sentiment analysis of the collected messages. This analysis evaluates the chatbot's ability to recognize and appropriately respond to positive, negative, or neutral expressions. Positive emotions accounted for the largest proportion (39%), followed by negative (22%), slightly positive (18%), neutral (11%), slightly negative (9%), and very small proportions of clearly positive (1%) and clearly negative (0%) emotions. This distribution reflects the overall emotional patterns observed during user interactions, without implying causal changes over time.

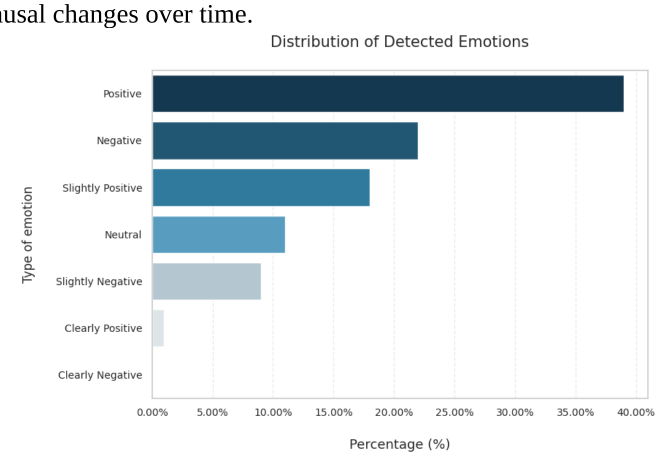

Fig. 4. Distribution of detected emotions in user interactions during the 30-day pilot study (n = 30).

To evaluate the performance of the sentiment classification component, a subset of 200 manually annotated messages was used. The classification performance was assessed using standard metrics, including accuracy, precision, recall, and F1-score.

Accuracy is defined as:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

Precision is defined as:

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

Recall is defined as:

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

F1-score is defined as:

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (4)$$

Where TP represents true positives, TN true negatives, FP false positives, and FN false negatives.

Based on the evaluation, the following values were obtained: TP = 95, TN = 90, FP = 10, and FN = 5. The sentiment classification model achieved an accuracy of 92.5%, a precision of 90.5%, a recall of 95.0%, and an F1-score of 92.7%. These results indicate a high level of performance, particularly in recall, suggesting that the model is well-suited for detecting negative emotional expressions.

C. *Difference in Average PHQ-9 score (Pre–Post)*

Fig. 5 shows the comparative analysis of average PHQ-9 scores before and after the interaction period. The mean pretest score was 13.2 (SD = 6.96), corresponding to moderate depressive symptoms, while the mean posttest score decreased to 6.1 (SD = 2.33), indicating mild symptoms. This represents a 53.79% reduction. A paired t-test indicated a statistically significant reduction in depressive symptoms after the intervention (t = 4.45, p = 0.0016). The effect size was large (Cohen's d = 1.41), suggesting a substantial magnitude of change within the sample.

These results suggest a positive trend in symptom reduction during the pilot phase. However, given the sample size (n = 30) and the absence of a control group, these findings should be interpreted as preliminary and do not establish clinical effectiveness.

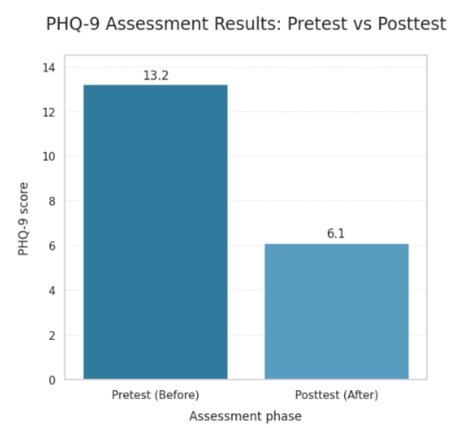

Fig. 5. Average PHQ-9 scores before and after the 30-day intervention (n = 30).

D. *User Usability and Satisfaction*

Fig. 6 summarizes post-test usability and satisfaction results obtained using the System Usability Scale (SUS) and additional questionnaires assessing coherence and empathy. The chatbot achieved an average SUS score of 90%, classified as Excellent. Perceived empathy reached 86.7%, response coherence reached 86.6%, and the retention rate was 70%, meeting the expected thresholds. These results suggest that users perceived the chatbot as an intuitive, empathetic, and useful tool for emotional support.

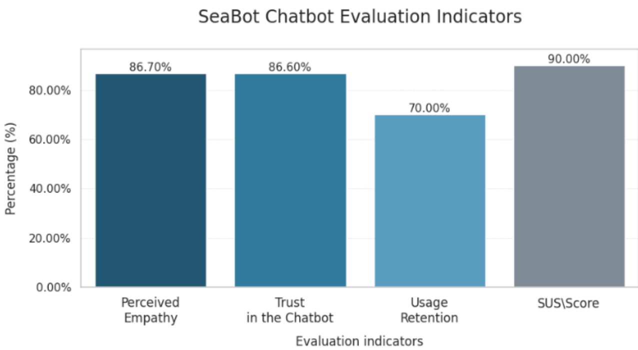

Fig. 6. SeaBot evaluation indicators, including perceived empathy, trust, retention, and usability (SUS score).

E. Chatbot Response Time

The average response time recorded during interactions was 2.32 seconds, remaining within the expected threshold of ≤ 3 seconds. This performance is appropriate for generative-model systems such as GPT-4, where latency depends on model processing and the infrastructure used in the pilot environment. Overall, the chatbot demonstrated stable response times that did not significantly affect user experience and were consistent with acceptable performance for real-time conversational systems.

VII. DISCUSSION

The results of this study suggest that the SeaBot chatbot is a feasible and well-accepted tool for providing emotional support among university students. The observed reduction in PHQ-9 scores—from a mean of 13.2 (moderate depression) to 6.1 (low depression), representing a 53.79% decrease—indicates a substantial improvement in participants' emotional well-being during the intervention period.

The high usability score (SUS ≈ 90%) further supports the system's potential for real-world adoption. A SUS score within this range is typically associated with excellent usability, suggesting that users perceived the system as intuitive and easy to interact with. In addition, the retention rate indicates that users maintained engagement over time, which is a critical factor in the effectiveness of digital mental health interventions.

When compared with previous studies, important differences can be observed. Kang and Hong [4] reported high user satisfaction in a GPT-4–based chatbot, with strong scores in positivity and support (mean 9.0/10), empathy (mean 8.7/10), and active listening (mean 8.0/10), along with statistically significant differences compared to other chatbot systems ($F = 3.27$, $p = .047$; $F = 12.94$, $p < .001$). However, their evaluation focused primarily on user experience metrics, identifying limitations in personalization and response quality, and did not include standardized clinical measures such as PHQ-9. In contrast, the present study not only achieved high usability and perceived empathy but also incorporated measurable changes in depressive symptoms, providing a more integrated evaluation of both technical and psychological dimensions.

Similarly, He et al. [13] conducted a randomized controlled trial using a CBT-based chatbot (XiaoE), reporting statistically significant reductions in depressive symptoms with moderate effect sizes ($d = 0.51$ at post-test and $d = 0.31$ at follow-up), as well as improvements in therapeutic alliance and acceptability. However, no significant differences were found in usability between intervention groups. Compared to these findings, the present study reports a larger observed reduction in PHQ-9 scores and higher usability levels, although it lacks the methodological rigor of a controlled experimental design. Therefore, while the magnitude of improvement appears greater, it should be interpreted cautiously due to the exploratory nature of the study.

From a technical perspective, the system's performance may be explained by the capabilities of GPT-4 in generating coherent, context-aware, and emotionally adaptive responses. These characteristics likely contributed to the high perceived empathy, conversational coherence, and sustained engagement observed among users. The integration of sentiment analysis and structured clinical assessment (PHQ-9) further supports a hybrid approach, combining generative AI with standardized evaluation tools.

From a practical standpoint, the system may serve as a complementary tool within university mental health services, particularly for early-stage emotional support, monitoring, and prevention. From a theoretical perspective, this study contributes to the growing body of research on AI-based mental health interventions by demonstrating how large language models can be combined with clinical screening instruments to provide both supportive interaction and measurable outcomes.

Overall, while the results are promising, further research with larger samples and controlled experimental designs is required to confirm these findings and establish clinical effectiveness.

VIII. LIMITATIONS

This study presents several limitations that should be considered when interpreting the results. First, the sample size was relatively small ($n = 30$), which limits the generalizability of the findings to broader populations.

Second, the study employed a quasi-experimental pretest–posttest design without a control group, which restricts the ability to establish causal relationships between chatbot use and observed changes in depressive symptoms.

Third, the intervention period was limited to one month, which may not be sufficient to evaluate long-term effects or sustained behavioral changes.

Additionally, the sentiment analysis component was validated using a subset of manually annotated messages, which, although useful for preliminary evaluation, may not fully capture the variability of emotional expression in real-world contexts.

Finally, the study relied on self-reported measures such as the PHQ-9, which may be subject to response bias and individual interpretation.

IX. CONCLUSIONS

The findings of this study indicate that SeaBot—a chatbot built on advanced generative models and Natural Language Processing techniques— demonstrates technical feasibility and high user acceptance as a tool for providing initial emotional support to Peruvian university students. Interaction with the system showed an observed reduction in participants' emotional well-being indicators, reflected in a decrease in depressive symptom intensity during the usage period, although these results should be interpreted as preliminary. Furthermore, conversational-quality and user-experience metrics highlight the chatbot's ability to deliver coherent, empathetic, and contextually appropriate responses, reinforcing its potential as a supportive technological resource.

The analysis of usage behavior revealed patterns consistent with those observed in digital mental-health interventions, where engagement levels tend to be higher during the initial stages of interaction. Despite this, the system maintained consistent performance, offering fluid responses and a stable experience suggesting its potential to integrate into student routines that demand immediacy and accessibility. This technical stability supports its potential use as a complementary resource in settings where emotional support availability is limited.

Overall, the results position SeaBot as a promising alternative within the ecosystem of digital tools aimed at supporting university mental health. However, given the sample size and the absence of a control group, further studies with larger populations and controlled experimental designs are required to establish clinical effectiveness. In a context where mental health is gaining increasing relevance, solutions like SeaBot represent a meaningful step toward the responsible and effective integration of artificial intelligence technologies in emotional care.

X. RECOMMENDATIONS

Future work should focus on validating the chatbot with a larger and more diverse sample of university students to strengthen the generalizability of the results and assess system stability across different academic and emotional contexts. Additionally, incorporating supplementary clinical instruments—such as the GAD-7 or WHO-5—would enhance psychological assessment and increase the methodological rigor of the study.

Further developments should include advanced personalization mechanisms that allow the chatbot to adapt its responses to each user's communication style and emotional profile. The integration of psychoeducational modules and lightweight cognitive–behavioral strategies could further enhance the system's formative value, while maintaining appropriate ethical boundaries.

Finally, future improvements should focus on optimizing cloud infrastructure to reduce latency, as well as exploring alternative emotional-analysis models, including hybrid classifiers or affective embeddings, to increase accuracy in detecting complex emotional states. These advancements would contribute to the development of more scalable, efficient, and widely applicable digital solutions for educational environments.

REFERENCES

[1] World Health Organization (WHO). (2025). Mental health of adolescents. https://www.who.int/news-room/fact sheets/detail/adolescent-mental-health

[2] World Health Organization (WHO). (2025). Suicide. https://www.who.int/es/news-room/fact-sheets/detail/suicide

[3] Ministerio de Salud (MINSA). (2024). Sala situacional de salud mental. Gobierno del Perú. https://app7.dge.gob.pe/maps/sala_salud_mental/

[4] 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

[5] Karkosz, S., Szymański, R., Sanna, K., & Michałowski, J. (2024). Effectiveness of a web-based and mobile therapy chatbot on anxiety and depressive symptoms in subclinical young adults: randomized controlled trial. JMIR Formative Research, 8. https://doi.org/10.2196/47960

[6] Ulrich, S., Lienhard, N., Künzli, H., & Kowatsch, T. (2024). A chatbot-delivered stress management coaching for students (MISHA app): pilot randomized controlled trial. *JMIR mHealth and uHealth*, *12*, e54945. https://doi.org/10.2196/54945

[7] Potts, C., et al. (2023). A multilingual digital mental health and well-being chatbot (ChatPal): pre-post multicenter intervention study. *Journal of Medical Internet Research*, *25*. https://doi.org/10.2196/43051

[8] Rathnayaka, P., et al. (2022). A mental health chatbot with cognitive skills for personalised behavioural activation and remote health monitoring. *Sensors*, *22*(10). https://doi.org/10.3390/s22103653

[9] Eltahawy, L., Essig, T., Myszkowski, N., & Trub, L. (2024). Can robots do therapy? Examining the efficacy of a CBT bot in comparison with other behavioral intervention technologies in alleviating mental health symptoms. *Computers in Human Behavior*, *2*(1), 100035. https://doi.org/10.1016/j.chbah.2023.100035

[10] Kang, B., & Hong, M. (2024). Digital interventions for reducing loneliness and depression in Korean college students: a mixed-methods evaluation. *JMIR Formative Research*. https://doi.org/10.2196/58791

[11] Kaywan, P., Ahmed, K., Ibaida, A., Miao, Y., & Gu, B. (2023). Early detection of depression using a conversational AI bot: a nonclinical trial. *PLoS ONE*, *18*(2 February). https://doi.org/10.1371/journal.pone.0279743

[12] Maciejewski, J., & Smoktunowicz, E. (2023). Low-effort internet intervention to reduce students' stress delivered with Meta's Messenger chatbot (Stressbot): a randomized controlled trial. *Internet Interventions*, *33*. https://doi.org/10.1016/j.invent.2023.100653

[13] He, Y., et al. (2022). Mental health chatbot for young adults with depressive symptoms during the COVID-19 pandemic: single-blind, three-arm randomized controlled trial. *Journal of Medical Internet Research*, *24*(11). https://doi.org/10.2196/40719

[14] Goyal, S., Dutta, R. R., Dev, S., Raju, K. N., & Bhatt, M. W. (2025). MindLift: AI-powered mental health assessment for students. *Neuroscience Informatics*, *5*(2). https://doi.org/10.1016/j.neuri.2025.100208

[15] Obadinma, S., et al. (2025). The FAIIR conversational AI agent assistant for youth mental health service provision. *npj Digital Medicine*, *8*(1). https://doi.org/10.1038/s41746-025-01647-6

[16] Kim, M., et al. (2025). Therapeutic potential of social chatbots in alleviating loneliness and social anxiety: quasi-experimental mixed methods study. *Journal of Medical Internet Research*, *27*. https://doi.org/10.2196/65589

[17] Chin, H., et al. (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*. https://doi.org/10.2196/51712

[18] Chong, M. K., et al. (2024). A digital approach for addressing suicidal ideation and behaviors in youth mental health services: observational study. *Journal of Medical Internet Research*, *26*. https://doi.org/10.2196/60879

[19] Yeo, G. H., Loo, G., Oon, M., Pang, R., & Ho, D. (2023). A Digital Peer Support Platform to Translate Online Peer Support for Emerging Adult Mental Well-being: randomized Controlled Trial. *JMIR Mental Health*, *10*. https://doi.org/10.2196/43956

[20] Meyerhoff, J., et al. (2025). Automated digital safety planning interventions for young adults: qualitative study using online co-design methods. *JMIR Formative Research*, *9*. https://doi.org/10.2196/69602

[21] Niu, W., et al. (2025). Decoding vital variables in predicting different phases of suicide among young adults with childhood sexual abuse: a machine learning approach. *Translational Psychiatry*, *15*(1). https://doi.org/10.1038/s41398-025-03360-0

[22] Cliffe, B., Gore-Rodney, J., Linton, M. J., & Biddle, L. (2023). Developing Suicide Prevention Tools in the Context of Digital Peer Support: qualitative analysis of a workshop with multidisciplinary stakeholders. *JMIR Formative Research*, *7*(1). https://doi.org/10.2196/47178

[23] Rainbow, C., Tatnell, R., Blashki, G., Fuller-Tyszkiewicz, M., & Melvin, G. A. (2024). Digital safety plan effectiveness and use: Findings from a three-month longitudinal study. *Psychiatry Research*, *333*. https://doi.org/10.1016/j.psychres.2024.115748

[24] Tong, A. C. Y., Wong, K. T. Y., Chung, W. W. T., & Mak, W. W. S. (2025). Effectiveness of topic-based chatbots on mental health self-care and mental well-being: randomized controlled trial. *Journal of Medical Internet Research*, *27*(1). https://doi.org/10.2196/70436

[25] Weiner, L. S., et al. (2023). Engagement, acceptability, and effectiveness of the Vira digital behavior change platform: a randomized pilot trial comparing self-care and coach-supported versions in young adults at risk for depression and obesity (Preprint). *JMIR Mental Health*. https://doi.org/10.2196/51366

[26] Torok, M., et al. (2022). The effect of a therapeutic smartphone application on suicidal ideation in young adults: findings from a randomized controlled trial in Australia. *PLoS Medicine*, *19*(5). https://doi.org/10.1371/journal.pmed.1003978

[27] Litvin, S., et al. (2023). The impact of a gamified mobile mental health app (eQuoo) on resilience and mental health in a student population: large-scale randomized controlled trial. *JMIR Mental Health*, *10*. https://doi.org/10.2196/47285

[28] Wrightson-Hester, A. R., et al. (2023). An artificial therapist (Manage Your Life Online) to support the mental health of youth: co-design and case series. *JMIR Human Factors*, *10*. https://doi.org/10.2196/46849

[29] Park, D. Y., & Kim, H. (2023). Determinants of intentions to use digital mental healthcare content among university students, faculty, and staff: motivation, perceived usefulness, perceived ease of use, and parasocial interaction with AI chatbot. Sustainability (Switzerland), 15(1). https://doi.org/10.3390/su15010872

[30] 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). https://doi.org/10.1145/3485874

[31] Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med. 2001 Sep;16(9):606-13. doi: 10.1046/j.1525-1497.2001.016009606.x. PMID: 11556941; PMCID: PMC1495268.