Generative Artificial Intelligence Chatbot to Diagnose the Severity of Anxiety in Unsch Students Using the Hamilton Anxiety Scale (HAM-A), 2025

Hubner Janampa Patilla[1], Efraín Porras Flores[1], Edem Terraza Huamán[2], Christian Lezama Cuellar[2], Lissette Elvira Fernández Jerí[2], Richard Zapata Casaverde[2] and Yudith Meneses Conislla[2]
[1] Department of Mathematics and Physics, National University of San Cristóbal de Huamanga UII-FIMGC, Peru
[2] Department of Mathematics and Physics, National University of San Cristóbal de Huamanga, School of Systems Engineering, Peru.
E-mail: hubner.janampa@unsch.edu.pe, efrain.porras@unsch.edu.pe, edem.terraza@unsch.edu.pe, christian.lezama@unsch.edu.pe, lissette.fernandez@unsch.edu.pe, richard.zapata@unsch.edu.pe, yudith.meneses@unsch.edu.pe

Keywords: Artificial intelligence, chatbot, anxiety, HAM-a, Cohen's kappa, confusion matrix

Received: July 29, 2026

Anxiety is a prevalent mental health problem among university students, creating the need for scalable and reliable tools for screening, severity classification, and decision support. This pilot study aimed to design, develop, and preliminarily evaluate a generative artificial intelligence chatbot for screening anxiety severity among Engineering students at the Universidad Nacional de San Cristóbal de Huamanga, Peru, during the 2025-II academic semester, using the Hamilton Anxiety Rating Scale (HAM-A). A pilot sample of 50 students was selected through cluster sampling. The chatbot, developed with Flutter and Dart and integrated with a generative language model API, automated HAM-A administration, score calculation, severity classification, result storage, and personalized feedback generation. The system was preliminarily validated by comparing chatbot-generated severity classifications with those assigned by a clinical mental health expert. Agreement was assessed using Cohen's Kappa coefficient and a confusion matrix, while classification performance was analyzed using class-specific and multiclass metrics. The results showed 46 exact matches out of 50 cases, with an observed agreement of 92.0% and Cohen's Kappa of $\kappa = 0.8086$, indicating almost perfect concordance in this pilot sample. Overall accuracy was 0.9200, with a Weighted F1-score of 0.9032, Macro F1-score of 0.7560, and Balanced Accuracy of 0.7197. The chatbot showed strong performance in moderate anxiety classification and robust identification of severe anxiety cases, with no false negatives in the severe category. However, the mild category showed lower recall, indicating a tendency to classify some mild cases into higher severity levels. These findings suggest that the chatbot has preliminary potential as a first-level screening and decision-support tool for university mental health programs. Nevertheless, due to the pilot nature of the study and the limited sample size, the results should be interpreted cautiously and validated in larger, more diverse, and preferably multicenter samples. The chatbot should be considered a complementary support system, not a substitute for formal clinical diagnosis or professional psychological evaluation.

Povzetek: Študija predstavlja AI-klepetalnega robota za presejanje anksioznosti pri študentih, ki je pokazal visoko ujemanje s strokovno oceno in potencial za podporo univerzitetnim programom duševnega zdravja.

1 Introduction

Anxiety represents one of the most widespread and concerning mental health problems among the global university population. Recent studies have estimated that approximately one third of higher education students experience elevated anxiety symptoms, with a pooled prevalence close to 39% [1]. This phenomenon not only reflects individual challenges related to psychological well-being but also entails academic, social, and economic consequences, as anxiety can negatively affect academic performance, student retention, and overall health [2]. The university environment, characterized by life transitions, academic demands, and social changes, creates a context of vulnerability that has been further intensified by the COVID-19 pandemic [3].

In parallel, the global shortage of mental health resources—particularly in low- and middle-income countries—and the limited access to traditional assessment and care services have driven the search for scalable digital solutions [4]. In this context, generative artificial intelligence (AI) chatbots have emerged as promising tools, offering conversational support, 24/7 availability, and personalization capabilities at a substantially lower cost than traditional clinical

interventions [5]. However, despite the rapid evolution of chatbot architectures—from rule-based systems to large language model (LLM)-based approaches—only a small proportion of studies have reached the stage of rigorous clinical validation [6].

Within this framework, anxiety assessment using well-established clinical–psychological instruments, such as the Hamilton Anxiety Rating Scale (HAM-A), remains a valuable standard in both practice and research [7]. Nevertheless, its traditional application requires structured interviews and trained personnel, which limits large-scale deployment in university settings.

The present pilot study focuses on the development and implementation of a generative artificial intelligence chatbot designed to support anxiety screening and severity classification among students at the National University of San Cristóbal de Huamanga, located in Ayacucho, Peru, using the Hamilton Anxiety Rating Scale (HAM-A) as the reference psychometric instrument. The primary objective is to build a generative AI system capable of automatically classifying anxiety levels—mild, moderate, or severe—and to preliminarily evaluate the degree of concordance between the severity categories generated by the chatbot and those established through conventional clinical assessment conducted by a mental health expert. Given the pilot scope of the study and the sample size of 50 participants, the findings should be interpreted as preliminary evidence of feasibility and agreement, rather than as definitive evidence for broad institutional or clinical implementation.

To this end, system validation is carried out by estimating Cohen's Kappa coefficient, which quantifies the level of agreement between both evaluators beyond chance and determines the reliability of the chatbot as a screening support tool for psychological assessment in the university context.

This pilot study was guided by the following research questions:

RQ1. Can a generative AI chatbot be designed and implemented to administer the HAM-A scale, calculate total scores, classify anxiety severity, store results, and generate personalized feedback?

RQ2. What is the level of concordance between anxiety severity classifications generated by the chatbot and those assigned by a clinical mental health expert?

RQ3. What class-specific and overall classification performance does the chatbot achieve in distinguishing mild, moderate, and severe anxiety levels?

The general objective of the study was to design, develop, and preliminarily evaluate a generative AI chatbot for anxiety severity screening and classification using the HAM-A scale among university students.

The general objective of this study was to design, develop, and preliminarily evaluate a generative AI chatbot for anxiety severity screening and classification among university students using the HAM-A scale. Specifically, the study aimed to implement a chatbot-based system capable of administering the 14 HAM-A items and computing anxiety severity levels; to integrate generative AI–based feedback into the assessment workflow while maintaining the system as a screening and support tool; to evaluate the agreement between chatbot-generated classifications and expert clinical classifications using Cohen's Kappa coefficient; and to analyze class-specific and overall classification performance using a confusion matrix, precision, recall, specificity, F1-score, accuracy, Macro F1, Weighted F1, and Balanced Accuracy. Because this was an exploratory pilot study, no causal or predictive hypothesis was formulated; instead, the study focused on preliminary feasibility, classification performance, and agreement with expert judgment.

2 Literature review

Anxiety among university students is widely recognized as a global public health problem, with prevalence rates exceeding 30% across multiple cultural contexts [1]. This condition is associated with academic, social, and emotional consequences, including reduced academic performance, difficulties in student retention, and deterioration of overall well-being [2]. Factors such as academic transitions, performance pressure, economic instability, and the psychosocial changes inherent to this life stage substantially increase the risk of developing anxiety disorders [3]. In settings with limited availability of psychological services, this burden is further exacerbated, highlighting the need for innovative, accessible, and scalable strategies that support anxiety screening, severity classification, decision-making, and emotional guidance [4].

Generative artificial intelligence (AI)–based chatbots have emerged as promising tools for supporting mental health-related processes, particularly through personalized interventions and conversational interaction. Recent studies indicate that these systems may contribute to reducing symptoms of anxiety and depression, especially when they incorporate cognitive behavioral therapy principles, personalized feedback, and frequent user interaction [8-10]. Manole et al. [8] showed the potential of chatbot-based interventions for personalized anxiety management, while Zhang et al. [9] reported that generative AI mental health chatbots may function as therapeutic support tools for reducing mental health symptoms. Similarly, Mayor [10] synthesized evidence from reviews on chatbots and mental health, emphasizing their growing relevance but also the need for careful evaluation.

Controlled and pilot studies have also provided evidence regarding the feasibility, acceptability, and preliminary effectiveness of AI-based conversational systems in mental health. Heinz et al. [11] reported results from a randomized trial of a generative AI chatbot for mental health treatment, while Chen et al. [12] compared an AI chatbot with a nurse hotline in reducing anxiety and depression levels in the general population. In the

university context, Reyes-Portillo et al. [13] evaluated a generative AI-powered mental wellness chatbot for college students, reporting feasibility, acceptability, and improvements in well-being. These studies suggest that chatbot-based systems may support mental health care processes, although their use should remain complementary and subject to professional oversight [11-13].

Despite these promising findings, the literature also warns about important limitations and ethical risks. Generative AI chatbots may present difficulties in accurately recognizing clinical severity, managing vulnerable users, preventing overreliance, and ensuring safe responses in sensitive mental health contexts [14-17]. Head [14] discusses the broader impact of the AI revolution on mental health, including potential risks in crisis situations. Saraç et al. [15] highlight the limitations of AI chatbots as diagnostic tools, while Wang et al. [16] examine user attitudes and trust toward generative AI chatbots for social anxiety support. Casu et al. [17] further emphasize that although AI chatbots show effectiveness and feasibility, their implementation must consider clinical supervision, ethical safeguards, and clear boundaries regarding their non-diagnostic role. Therefore, the converging evidence underscores that these systems should be understood as complementary tools, not replacements for professional clinical evaluation [14-17].

The use of digital technologies in mental health, including self-assessment systems, internet-based interventions, mobile applications, and conversational agents, has grown steadily over the past decade. Evidence indicates that digital interventions can achieve effectiveness comparable to face-to-face care for mild and moderate mental health problems, particularly when they are structured, evidence-based, and appropriately monitored [18]. Within this framework, artificial intelligence–based chatbots have emerged as a promising alternative by offering continuous interaction, anonymity, immediate feedback, and a perceived sense of companionship [19]. In parallel, AI has become a relevant technology for clinical decision support, enabling the processing of large volumes of data and the identification of patterns useful for assessment, monitoring, and health-related decision-making [20]. This convergence of human and artificial intelligence contributes to high-performance medicine, where computational tools support diagnostic precision, personalization, and professional decision-making without replacing clinicians [21]. Accordingly, digital psychiatry increasingly integrates AI-driven applications and chatbots to support assessment, monitoring, and intervention in mental health, expanding access and complementing traditional clinical services [22].

The literature identifies three main generations of mental health chatbots: rule-based systems, models grounded in supervised learning, and conversational assistants powered by large language models (LLMs) [23]. This third generation, based on generative artificial intelligence, represents a qualitative advancement by enabling semantic understanding, empathetic response generation, and contextual adaptation [5]. Previous studies with conversational agents such as Tess and Woebot have reported reductions in anxiety and depression symptoms through automated psychological support and cognitive behavioral therapy–based interaction [23,24]. In addition, agreement and evaluation metrics such as weighted Kappa have been proposed as useful tools for assessing consistency between observers or classification systems, especially when categorical outcomes are involved [25]. Nevertheless, the literature consistently emphasizes the need for rigorous psychometric validation, ethical safeguards, and clinically responsible implementation to ensure that AI systems are safe and useful in mental health contexts [6]. Overall, AI-based chatbots show promising results in improving well-being and reducing anxiety and depression among university students, particularly when they integrate cognitive behavioral therapy, frequent interactions, and personalized interventions [32].

For anxiety assessment, structured clinical and psychometric instruments remain the primary methodological reference. Among them, the Hamilton Anxiety Rating Scale (HAM-A) stands out for its consistency, sensitivity, and extensive use in research and clinical practice since its development [26]. Evidence has supported the reliability, validity, and sensitivity to change of the HAM-A across diverse populations and clinical conditions [7]. Complementarily, Kessler et al. [27] validated brief screening scales for nonspecific psychological distress, demonstrating their efficiency in identifying symptoms related to anxiety and depression. These standardized instruments support the development of structured clinical systems, digital screening tools, and classification-based support platforms. In this context, artificial intelligence–based conversational agents have emerged as promising tools for facilitating the expression of emotional difficulties and providing scalable access to initial guidance [28].

A central component in the validation of automated tools such as chatbots is concordance with established instruments and expert judgment. Cohen's Kappa coefficient (κ) is widely used to estimate agreement beyond chance between categorical assessment methods [29]. This metric enables quantification of the concordance between the classifications generated by an automated system and those obtained through conventional application of a validated instrument by a clinical expert. In this regard, digital health and methodological literature recommends Kappa-based agreement analysis as a useful approach for evaluating automated systems and classification procedures, particularly when categorical severity levels are compared [25,29]. To better contextualize the proposed system within the current state of the art, Table 1 summarizes representative studies on artificial intelligence–based chatbots and digital tools for mental health assessment or intervention. The comparison includes the target population, computational approach, evaluation metrics, main findings, and reported limitations. This analysis helps identify relevant gaps, including limited clinical

validation, emphasis on user engagement rather than expert agreement, lack of standardized psychometric integration, and insufficient use of generative AI for structured anxiety severity classification.

Table 1: Comparative summary of related studies on AI-based mental health chatbots and digital assessment tools.

Authors / Year	Dataset or Population	AI / ML Method	Evaluation Metrics	Main Results	Limitations
Inkster et al. [19]	Real-world users of Wysa	Conversational AI chatbot	Symptom improvement, engagement, user feedback	Reported improvement in depressive symptoms and positive user acceptance	Focused mainly on engagement and symptom change; limited expert-based diagnostic validation
Fulmer et al. [23]	Adults using Tess chatbot	Psychological AI chatbot	Anxiety and depression symptom reduction	Significant reduction in anxiety and depression symptoms	Evaluated therapeutic effect rather than agreement with clinical expert classification
Fitzpatrick et al. [24]	Young adults with anxiety and depression symptoms	Fully automated CBT conversational agent	PHQ-9, GAD-7, symptom change	Short-term reduction in anxiety and depression symptoms	Limited follow-up and no HAM-A-based severity classification
Manole et al. [8]	Users receiving chatbot-based anxiety support	AI chatbot intervention	Anxiety management outcomes	Demonstrated potential for personalized anxiety support	Limited formal clinical validation and limited comparison with expert judgment
Zhang et al. [9]	Studies on generative AI mental health chatbots	Systematic review / meta-analysis	Pooled effects on mental health symptoms	Found potential of generative AI chatbots to reduce mental health symptoms	Heterogeneity of studies and limited standardization of validation procedures
Heinz et al. [11]	Participants in randomized trial	Generative AI chatbot	Clinical symptom outcomes	Reported positive outcomes in mental health treatment context	Focused on treatment outcomes rather than structured screening concordance
Chen et al. [12]	General population sample	AI chatbot compared with nurse hotline	Anxiety and depression reduction	Showed potential to reduce anxiety and depression levels	Pilot study; limited generalizability
Reyes-Portillo et al. [13]	College students	Generative AI-powered mental wellness chatbot	Feasibility, acceptability, well-being outcomes	Reported feasibility and acceptability in college students	Did not focus on HAM-A scoring or expert agreement
Casu et al. [17]	Scoping review of AI mental health chatbots	Review of chatbot applications	Effectiveness, feasibility, applications	Identified promising applications in mental health support	Highlighted ethical, methodological, and clinical validation limitations

From a software design perspective, the development of digital health chatbots also requires clear architectural organization, maintainability, and separation of responsibilities. Layered software architecture provides a structured foundation for organizing user interface components, application logic, data persistence, and external service integration [30]. Likewise, enterprise application architecture principles emphasize modularity, separation of concerns, and clear communication between layers, which are essential for developing scalable and auditable systems [31]. These principles are particularly relevant in the design of AI-assisted mental health tools, where traceability, reliability, and responsible data handling are required to support future validation and institutional adoption.

3 Materials and methods

3.1 Chatbot software architecture

The software architecture of the generative artificial intelligence chatbot for anxiety assessment using the Hamilton Anxiety Rating Scale (HAM-A) is structured according to a classical three-layer model: presentation, application/domain logic, and data–integration. This layered approach aligns with established best practices in the design of distributed systems and enterprise applications described in the software architecture literature [30,31], and is particularly appropriate for digital health solutions that require structural clarity, maintainability, and traceability [4,6].

Figure 1 illustrates the presentation layer, which is fully implemented in Flutter and is responsible for direct interaction with the student. It includes the *HomeScreen*, *HamaChatScreen*, *ReportListScreen*, and *ReportDetailScreen*, which together constitute the complete user workflow. The *HomeScreen* serves as the entry point, providing access to the assessment and the report history, while integrating institutional visual

elements and welcome messages. The *HamaChatScreen* manages the administration of the HAM-A scale by presenting items sequentially, displaying a progress indicator, and offering response options in the form of selectable "chips" that function as single-choice buttons, enhanced with visual emphasis and animations to support user focus during response selection. The *ReportListScreen* displays the set of stored assessments, and the *ReportDetailScreen* presents detailed information for a specific session, including individual items, responses, total score, severity level, and the AI-generated feedback. This layer interacts exclusively with models and services from the logic layer and does not directly handle data persistence or HTTP requests, in accordance with the principle of separation of concerns [31].

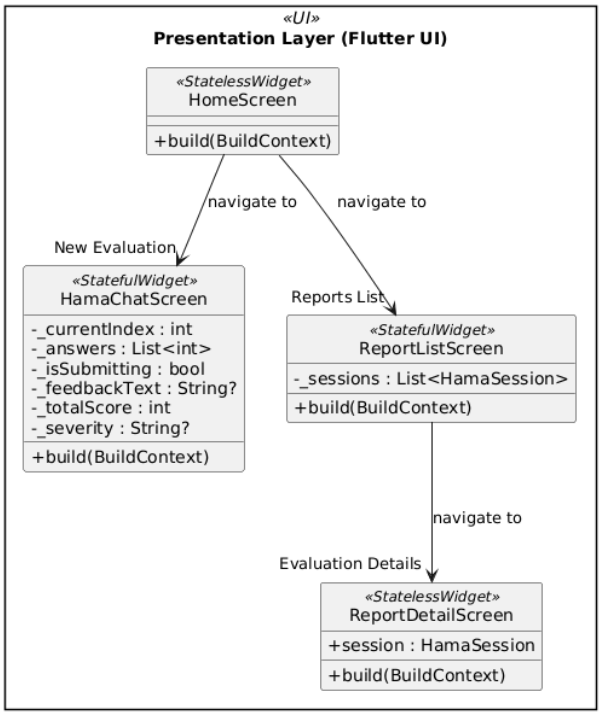


Figure 1: Presentation layer for the generative AI chatbot – HAM-a

Figure 2 illustrates the application and domain logic layer, which encapsulates the system-specific rules, the representation of core entities, and the orchestration of the assessment workflow. This layer includes the *HamaSession* model, the *HamaConstants* component, and the *StorageService* and *OpenAIService* services. *HamaSession* represents each assessment through a unique identifier, date, responses to the 14 HAM-A items, total score, severity category, and the generated feedback text. *HamaConstants* groups the item statements, response option labels, and the *classifySeverity* function, which translates the total score into severity levels such as mild, moderate, or severe anxiety, in accordance with the clinical logic of the original instrument [26,7]. *StorageService* abstracts the logic related to session management, including saving, listing, deleting, and exporting sessions to CSV format, while *OpenAIService* centralizes prompt construction, consumption of the generative API, and processing of natural language response. In this way, the logic layer acts as an intermediary between the user interface and the underlying technologies and constitutes the core where the rules of the domain "anxiety assessment using HAM-A supported by generative AI" are implemented [5,6].

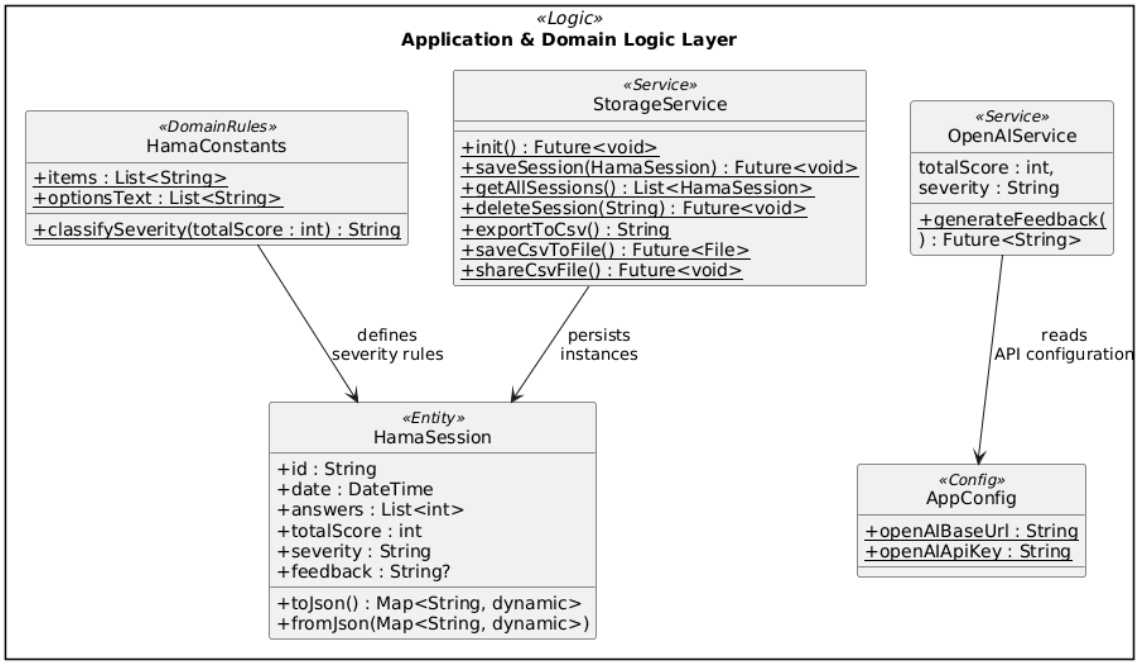


Figure 2: Architecture of the application and domain logic layer for the HAM-a generative AI chatbot

Figure 3 presents the data and integration layer, which is responsible for interaction with local storage mechanisms and external services. For local storage, Hive is used as a lightweight key–value database engine, integrated through StorageService, and supported by path_provider for accessing file system paths and share_plus for exporting and sharing CSV files containing the results. This structure enables assessment sessions to be persistently stored for subsequent analysis, methodological auditing, and the calculation of agreement measures such as Cohen's Kappa coefficient in specialized statistical environments [29]. Assessment sessions were stored locally in Hive, including the assessment date, HAM-A responses, total score, severity classification, and AI-generated feedback. The *StorageService* component managed report creation, retrieval, deletion, and export in CSV/PDF formats. To protect privacy, no personal identifiers were included, access was restricted to the local application environment, and encryption mechanisms provided by Hive were used.

At the level of external integration, *OpenAIService* communicates with the OpenAI API via the http library, using the configuration defined in *AppConfig* (base URL and API key). The construction of an ethical system prompt and a contextualized user prompt (including total score and severity level) allows the language model to generate empathetic, concise, and understandable feedback, in line with recent recommendations for the responsible design of mental health chatbots [19,5]. Strict adherence to the UI → application logic → data/AI flow, together with the absence of inverse dependencies, ensures a clear layered architecture in which each level has well-defined and testable responsibilities—an essential conditions for evaluation and eventual adoption in clinical contexts [30,4].

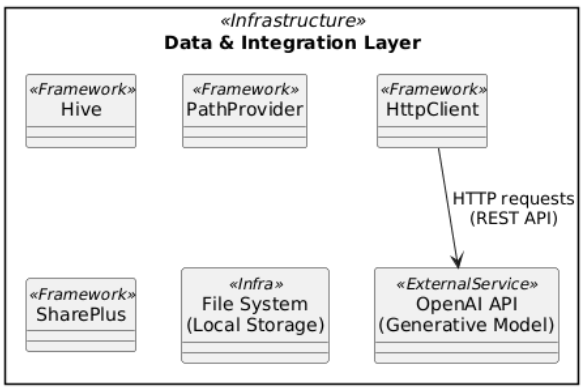


Figure 3: Architecture of the data and integration layer for a generative AI–based HAM-a chatbot

Figure 4 presents the simplified three-layer diagram representing the structural organization of the generative chatbot based on the HAM-A scale. The Presentation Layer groups all screens and interface components built in Flutter, which are responsible for direct interaction with the student, navigation between modules, and visualization of assessment results. The Application and Domain Logic Layer contain the functional core of the system. This layer manages the data model (*HamaSession*), the constants and rules associated with the HAM-A scale, score processing, anxiety level classification, and the services responsible for storing assessments or requesting AI-generated feedback. It operates as an intermediary between the user interface and the underlying infrastructure.

The Data and Integration Layer encompass the components responsible for local storage, file generation and export, HTTP communication, and interaction with the external OpenAI API. This layer ensures data persistence, secure data exchange, and connectivity with the generative model. The vertical structure of the diagram emphasizes separation of concerns: the interface invokes functions from the logic layer, which in turn delegates storage or external communication operations to the lower layer.

The HAM-A scale was administered through the chatbot using a conversational interface, in which the 14 items were presented sequentially and students selected their responses through interactive options. Each response was internally associated with a numerical value according to the scoring logic of the scale. After all items were completed, the system automatically calculated the total score and processed it through the classifySeverity() function, assigning the corresponding severity category: mild, moderate, or severe. This classification was performed deterministically by the application logic, while the generative language model was used only to generate personalized feedback.

Personalized feedback was generated by a generative AI model using the HAM-A total score and severity category calculated by the system. The process was guided by a structured prompt containing the score, severity level, and ethical instructions, ensuring that the feedback provided interpretation, self-care recommendations, and a disclaimer that the result was for screening purposes, not clinical diagnosis.

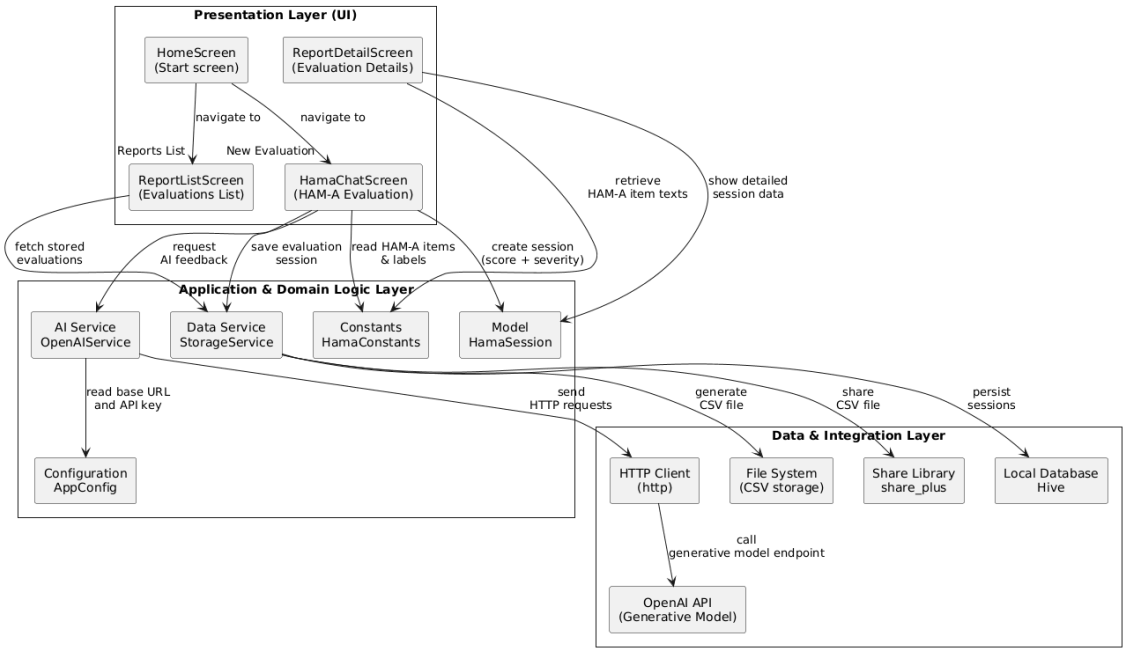


Figure 4: Simplified three-layer software architecture and data flow of the HAM-a generative AI chatbot

Figure 5 presents the sequence diagram corresponding to the anxiety assessment process implemented in the chatbot-based system using the Hamilton Anxiety Rating Scale (HAM-A). The diagram describes, in a chronological and structured manner, the dynamic interaction among the main participants of the system: the student, the user interface screens, the application logic layer, the local storage service, and the external artificial intelligence service. This representation is useful because it allows the complete workflow of the assessment to be understood from the moment the user initiates a new evaluation until the results are generated, stored, and displayed.

The process begins when the student interacts with the *HomeScreen* and selects the option "New assessment". This action represents the starting point of the evaluation flow and triggers the navigation process toward the *HamaChatScreen,* which is responsible for administering the HAM-A questionnaire through a chatbot-style interface. Once the student enters this screen, the system begins with the sequential presentation of the HAM-A items. The use of a chatbot interface allows the questionnaire to be administered in a more interactive and guided manner, reducing the need for the user to understand the complete structure of the scale before answering.

The diagram includes a loop fragment that represents the repeated interaction carried out for each item of the HAM-A instrument. Since the HAM-A scale consists of 14 items, this loop indicates that the system repeats the same response registration process until all items have been answered. For each question, the student selects an option using the interface components, such as chips or buttons. Each selected answer is then registered internally by the *HamaChatScreen* in the answers[index] structure. This temporary data structure stores the response associated with each item, ensuring that no answer is lost during the assessment process and that all responses remain available for score calculation once the questionnaire is completed.

After the student has answered all HAM-A items, the system activates the internal scoring logic. At this stage, the *HamaChatScreen* calculates the total score by summing the numerical values assigned to each individual response. This operation is represented in the diagram as the calculation of the totalScore. Immediately after obtaining the total score, the system determines the severity level through the *classifySeverity()* function. This function classifies the student's anxiety level according to the predefined interpretation thresholds of the HAM-A scale. Therefore, this stage transforms the raw answers provided by the student into clinically meaningful information, such as the total anxiety score and its corresponding severity category.

Once the score and severity level have been calculated, the system proceeds to generate personalized feedback. For this purpose, HamaChatScreen sends a request to the *OpenAIService* through the generateFeedback(totalScore, severity) operation. This service acts as an intermediary between the mobile application and the external OpenAI API. The *OpenAIService* prepares the request by including the total score, the severity classification, and the prompt instructions required to guide the generation of feedback. The diagram shows that this information is transmitted to the OpenAI API through an HTTP POST request containing the model and prompt parameters.

The OpenAI API processes the request and returns a JSON response containing the generated feedback. This response is received by the *OpenAIService*, which extracts and formats the feedback text before returning it to the *HamaChatScreen*. The purpose of this feedback is to provide the student with an explanatory and supportive interpretation of the result. In this sense, the use of artificial intelligence contributes to the personalization of the assessment output, since the feedback is not limited to displaying only the numerical score or severity label, but also includes a contextualized message that can help the student understand the meaning of the result within a university setting.

After receiving feedback, the system creates a HamaSession object. This object encapsulates the complete information generated during the assessment process, including the session identifier, the date of the evaluation, the list of individual answers, the total score, the severity classification, and the AI-generated feedback. The creation of this object is important because it organizes the assessment data into a structured unit that can be stored, retrieved, and later used by other modules of the application, such as the report or history module.

The *HamaSession* object is then sent to the *StorageService* through the *saveSession(HamaSession)* operation. The *StorageService* is responsible for managing the persistence layer of the application and communicating with the local Hive database. In the diagram, this interaction is represented by the *put(id, jsonString)* operation, where the session information is converted into a JSON string and stored using the unique session identifier as a key. Hive confirms the successful execution of the write operation by returning an OK response to the *StorageService.* Subsequently, the *StorageService* returns confirmation to the *HamaChatScreen,* indicating that the assessment session has been stored.

Once these operations have been completed, the *HamaChatScreen* displays the result card to the student. This result includes the total score obtained, the severity level identified by the system, and the personalized feedback generated through the artificial intelligence service. This step closes the assessment cycle and provides the student with immediate access to the interpretation of the results.

Overall, the sequence diagram demonstrates how the system coordinates user interaction, questionnaire administration, score computation, AI-based feedback generation, and local data persistence within a single integrated workflow.

The diagram also shows a clear separation of responsibilities among the system components: the *HomeScreen* manages the initial navigation, the *HamaChatScreen* controls the assessment logic and user interaction, the *OpenAIService* manages the communication with the external artificial intelligence API, the *StorageService* handles data persistence, and Hive stores the assessment records locally. This modular organization improves the maintainability of the system and facilitates the future extension of its functionalities, such as adding new psychological instruments, modifying severity classification rules, or integrating additional reporting features. The final phase of the workflow illustrates how the assessment screen presents the student with a results card displaying the total score, the anxiety classification, and the feedback message generated by the artificial intelligence model. This stage marks the completion of the process and demonstrates the seamless integration across system layers—from the initial user interaction in the interface, through local processing, to communication with external AI services.

The sequence diagram highlights the coherence and logical order of the system, as well as the appropriate separation of responsibilities among the user interface, the application logic layer, and the storage and generative AI services.

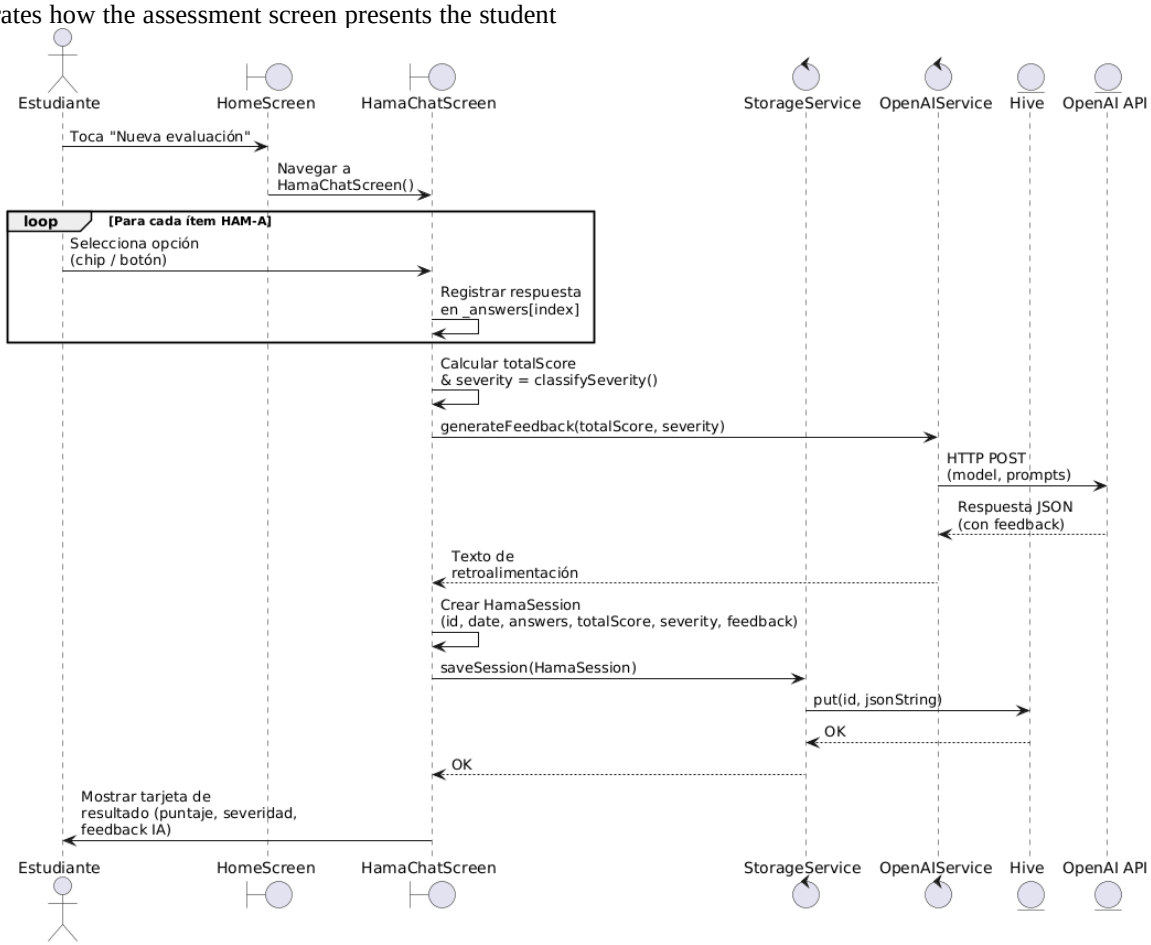


Figure 5: Sequence diagram of the HAM-a–based anxiety severity assessment using a generative AI chatbot

3.2 HAM-a scoring and feedback generation algorithm

Figure 6 presents the pseudocode of the proposed algorithm. The chatbot first collects the responses to the 14 HAM-A items, computes the total score, determines the anxiety severity level according to the predefined thresholds, constructs the system and user prompts, and submits the request to the OpenAI GPT-4o-mini model. The generated response is then validated and post-processed before being presented to the participant. When the API is unavailable or returns an invalid response, a standardized fallback message is automatically generated to ensure continuity of the screening process.

3.3 Performance evaluation metrics

Accuracy was used to report the overall proportion of correct classifications. However, because the dataset showed class imbalance across anxiety severity categories, complementary metrics were included to provide a more rigorous evaluation. Macro F1 was considered to assign equal weight to each class, regardless of its frequency, while Weighted F1 accounted for the number of cases in

each category. Balanced Accuracy was also reported to average recall across classes and reduce the influence of the majority class. In addition, class-specific precision, recall, specificity, and F1-score were calculated using a one-vs-rest approach to identify performance differences among mild, moderate, and severe anxiety categories.

3.4 Replicability and technical availability

To support replicability, the manuscript describes the main technical components of the system, including the Flutter-based interface, the HAM-A scoring logic, the StorageService module for local storage, and the OpenAIService module for AI-generated feedback. For privacy and security reasons, API keys and sensitive configuration files are not publicly shared. The source code or a demonstration version of the chatbot may be made available upon reasonable request to the corresponding author for academic and research purposes, subject to ethical and institutional restrictions.

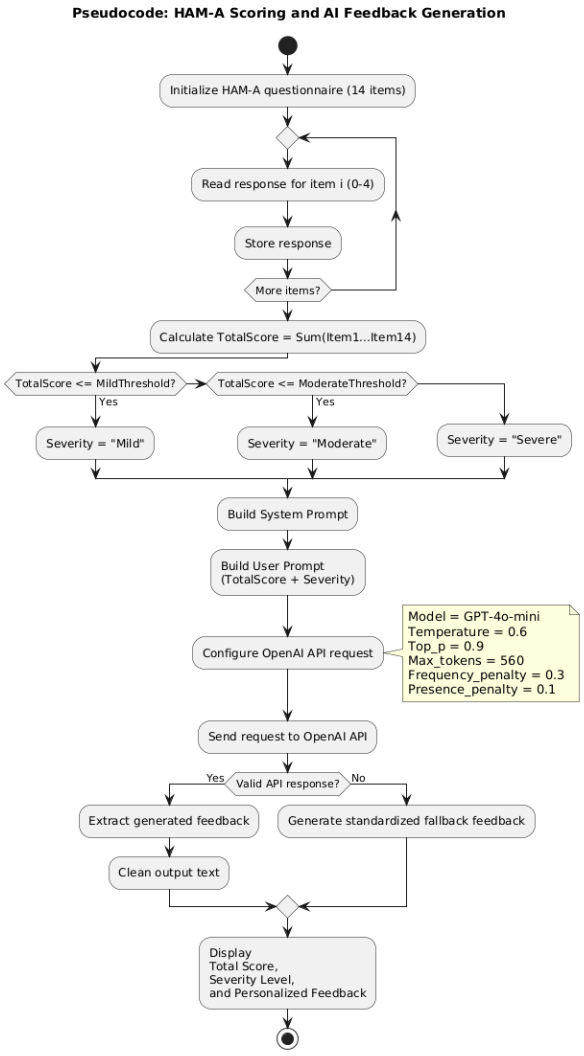


Figure 6: Algorithm 1. HAM-a scoring and AI-based feedback generation.

3.5 Ethical considerations

The study was conducted as a pilot evaluation and received approval from the Ethics Committee of the Universidad Nacional de San Cristóbal de Huamanga. All participants were informed about the purpose of the study, the voluntary nature of their participation, the confidentiality of their responses, and the non-diagnostic nature of the chatbot. Informed consent was obtained prior to participation. The chatbot was clearly presented as a screening and support tool for anxiety severity classification, not as a substitute for professional clinical diagnosis or psychological evaluation. The use of generative AI in mental health screening requires caution because AI-generated responses may be incomplete, overly general, or misinterpreted by users. In addition, classification errors may lead to false reassurance or unnecessary concern. Therefore, the chatbot was designed as a screening and decision-support tool, not as a diagnostic system. Human supervision remains necessary, especially for moderate or severe cases and for students reporting significant distress.

3.6 Generative AI model and prompt engineering

The chatbot was developed in Flutter and integrated with the commercial OpenAI API, using the GPT-4o-mini (gpt-4o-mini) large language model to generate personalized feedback based on the Hamilton Anxiety Rating Scale (HAM-A) results. No fine-tuning was performed; instead, the model's behavior was controlled through a prompt engineering strategy consisting of a system prompt, which defined the assistant's role, ethical constraints, and standardized interpretation of the HAM-A, and a user prompt, which dynamically incorporated the participant's total score and anxiety severity level to generate empathetic, structured, and non-diagnostic feedback. The API was configured with temperature = 0.6, top_p = 0.9, max_tokens = 560, frequency_penalty = 0.3, and presence_penalty = 0.1. In addition, the system implemented a fallback mechanism to provide standardized feedback whenever the API was unavailable or failed to return a valid response, ensuring consistent system operation.

4 Results

4.1 Cross-platform implementation of the HAM-a chatbot

The proposed solution was developed using the Flutter framework and the Dart programming language, enabling the implementation of a cross-platform application with a modular architecture, responsive interfaces, and consistent user experience. The graphical interfaces of the generative artificial intelligence chatbot based on the Hamilton Anxiety Rating Scale (HAM-A) are presented in Figures 6 to 18, illustrating the complete interaction workflow between the student and the chatbot, from the initial access

screen to the generation of the final report and personalized feedback.

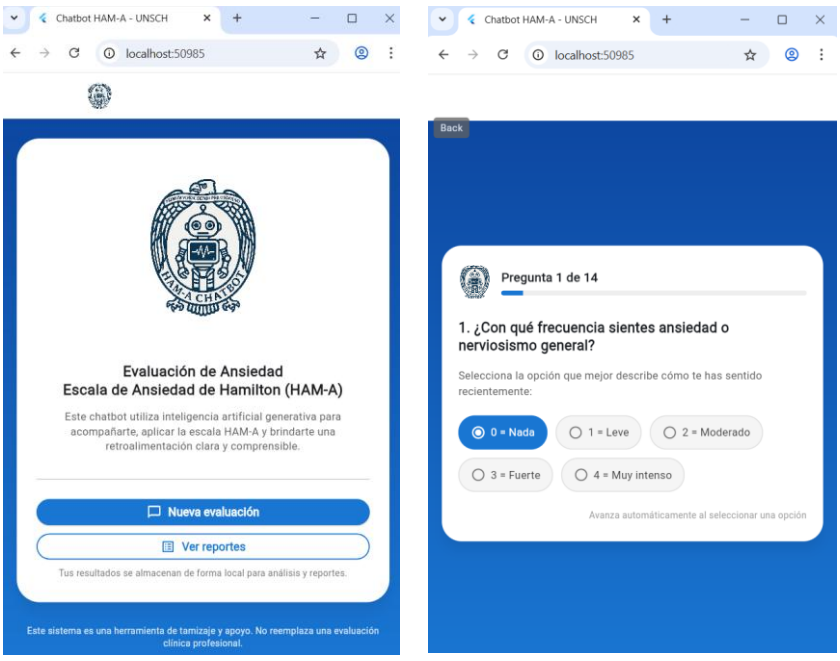
Figure 7 shows the home screen for initiating the anxiety assessment and the first HAM-A item related to anxiety or nervousness. Figures 8 to 13 present the sequential administration of the HAM-A items through the chatbot interface, including questions related to tension, fear, sleep difficulties, physical symptoms, digestive discomfort, tremors, concentration problems, restlessness, health-related worry, shortness of breath, weakness, fatigue, and repetitive negative thoughts. Figure 13 displays the final HAM-A item related to irritability and the first results screen, where the anxiety severity classification and automated feedback are presented.

Figure 14 presents the results screen, where the total HAM-A score, the automatically classified anxiety severity level—mild, moderate, or severe—and the personalized feedback generated by a large language model through the ChatGPT API are displayed. This feedback includes an interpretative summary of the results, practical recommendations, and ethical disclaimers, emphasizing that the system functions as a screening and support tool and does not replace professional clinical evaluation.

Figures 15 and 16 complement the assessment workflow by showing the stored evaluation summary, the item-level responses, and the detailed feedback associated with each completed HAM-A session. Figure 16 shows the evaluation summary screen with the assessment date, total score, and anxiety severity level, as well as a detailed screen displaying item-level responses. Figure 17 presents the remaining item-level responses and the full AI-generated feedback, including options to generate the PDF report and return to the previous screen.

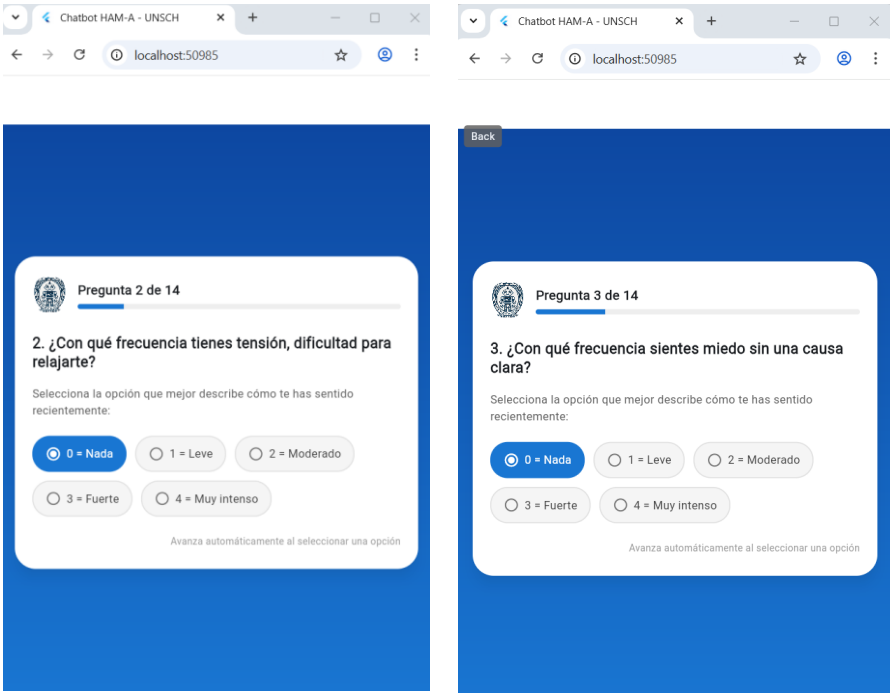
Figure 18 illustrates the automatically generated PDF report, which consolidates the HAM-A screening summary, item-level responses, severity classification, and generative AI feedback for documentation purposes.

Figure 19 presents the personalized feedback generated by the generative AI model, including a summary interpretation, practical recommendations, and ethical disclaimers. Together, these figures demonstrate the functional continuity of the application, the traceability of the assessment process, and the integration of screening, severity classification, feedback generation, and report production within a single chatbot-based system.



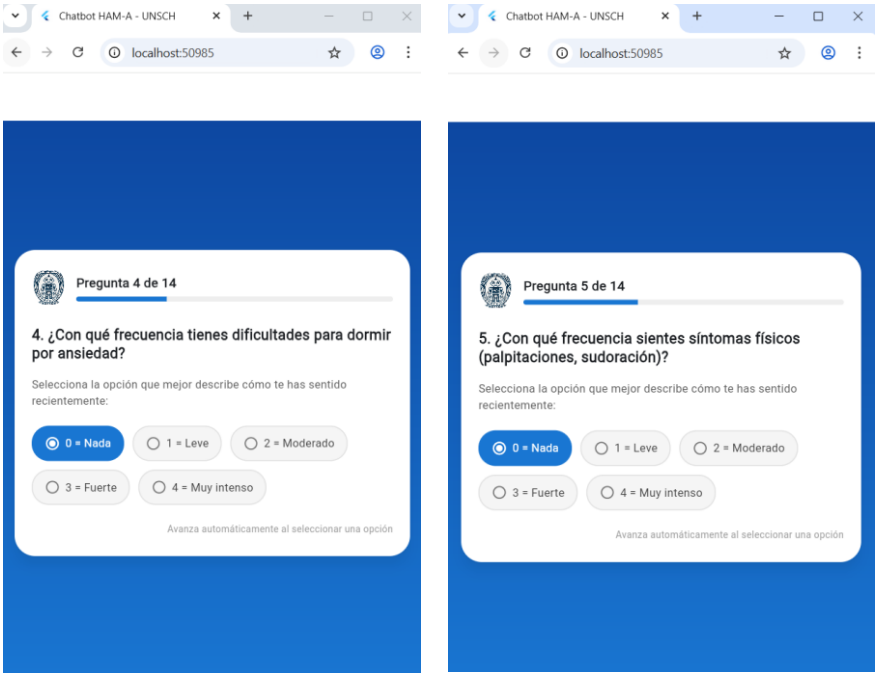
(a) (b)

Figure 7: Graphical user interface of the HAM-a generative AI chatbot: (a) home screen for initiating the anxiety assessment and (b) question on the frequency of anxiety or nervousness



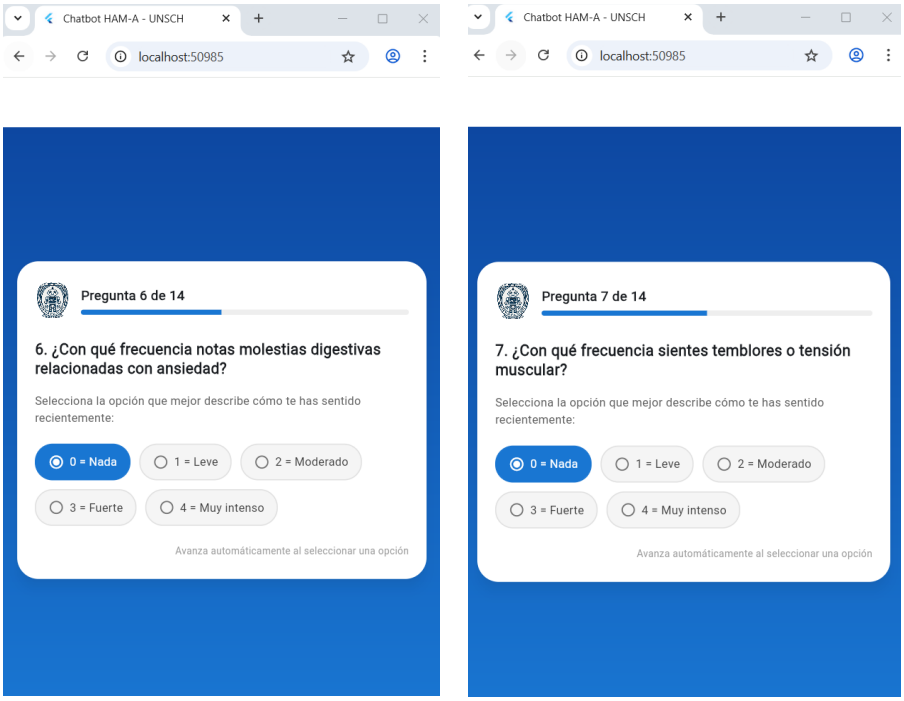
(a) (b)

Figure 8: Graphical user interface of the HAM-a generative AI chatbot: (a) question on tension and difficulty relaxing and (b) question on fear without an apparent cause



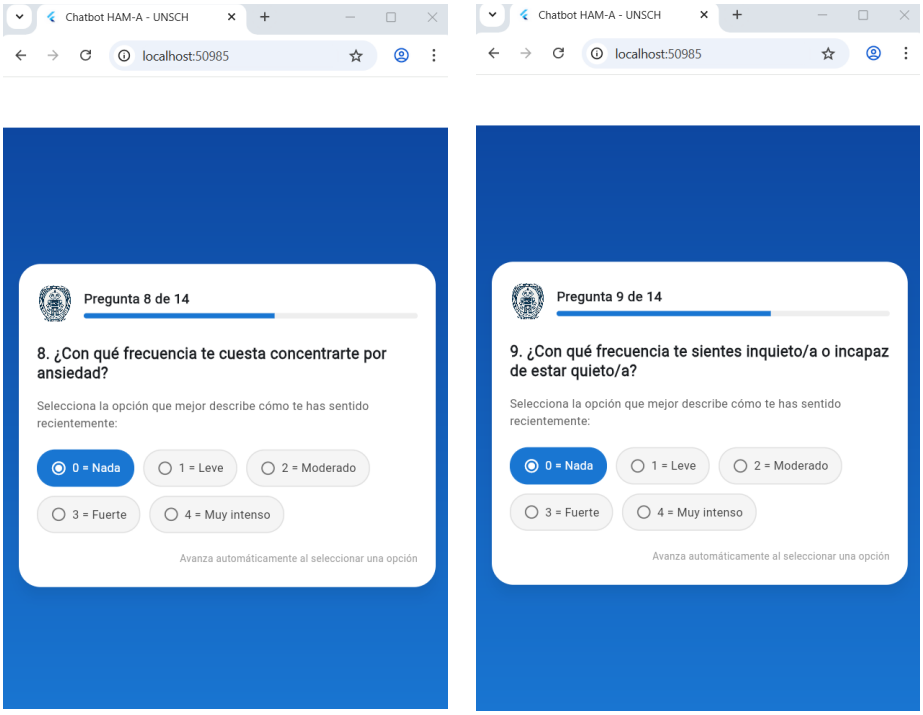
(a) (b)

Figure 9: Graphical user interface of the HAM-a generative AI chatbot: (a) question on sleep difficulties due to anxiety and (b) question on physical anxiety symptoms (palpitations and sweating)



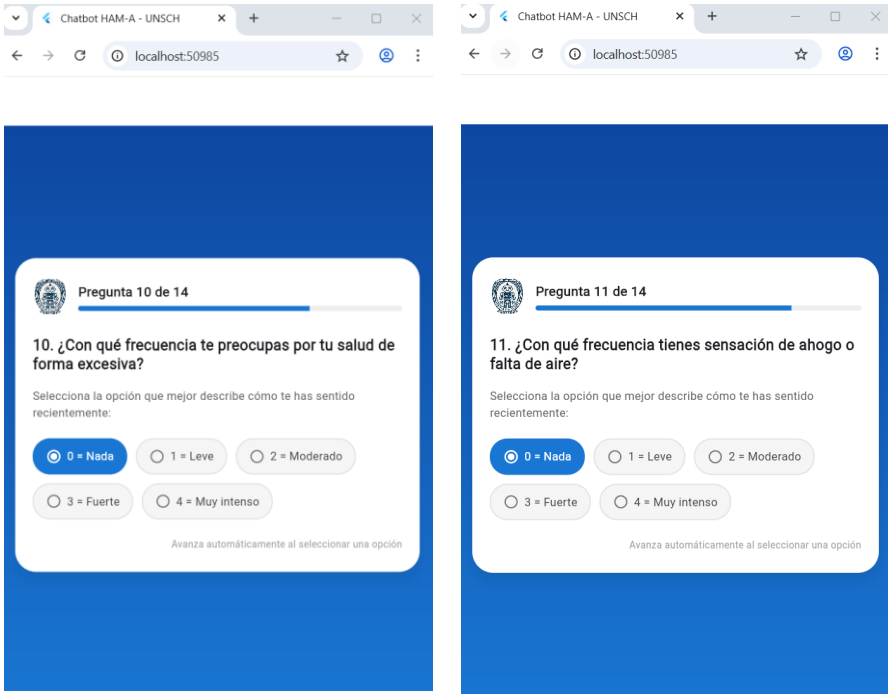
(a) (b)

Figure 10: Graphical user interface of the HAM-a generative AI chatbot: (a) digestive discomfort associated with anxiety and (b) question on tremors or muscular tension



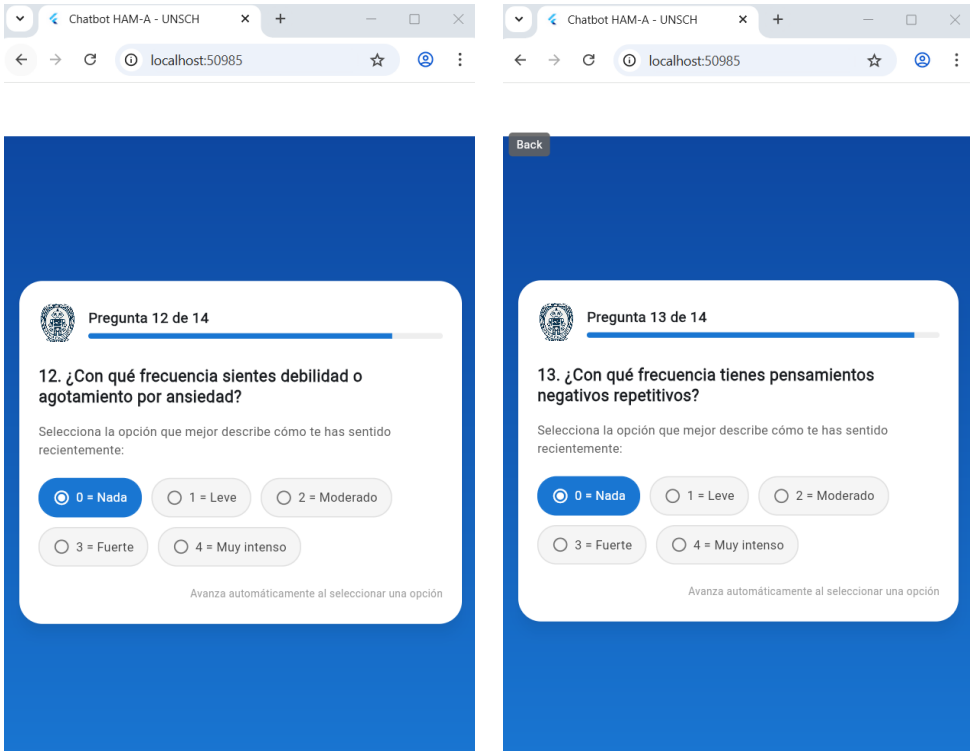
(a) (b)

Figure 11: Graphical user interface of the HAM-a generative AI chatbot: (a) question on difficulty concentrating due to anxiety and (b) question on restlessness or inability to remain calm



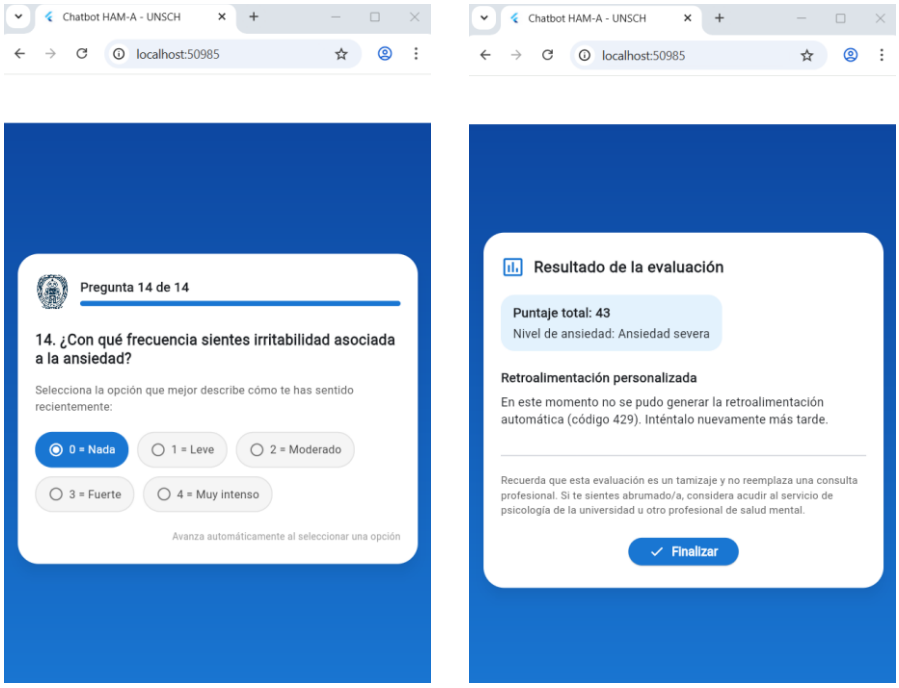
(a) (b)

Figure 12: Graphical user interface of the HAM-a generative AI chatbot: (a) question on excessive health-related worry and (b) question on sensations of shortness of breath or suffocation



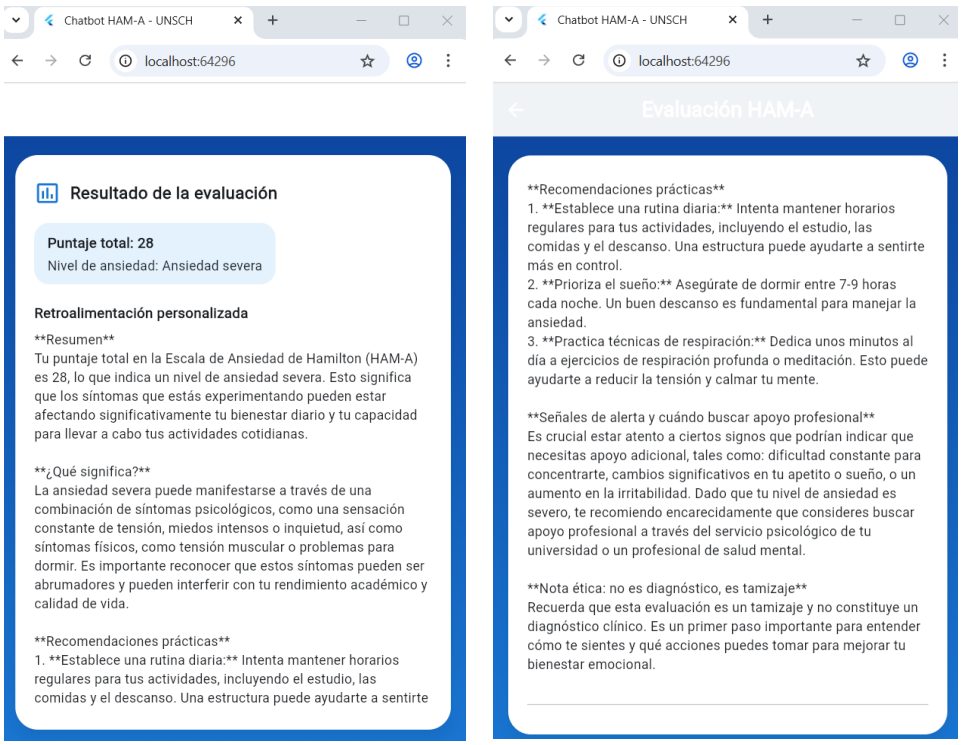
(a) (b)

Figure 13: Graphical user interface of the HAM-a generative AI chatbot: (a) question on feelings of weakness or fatigue due to anxiety and (b) question on repetitive negative thoughts



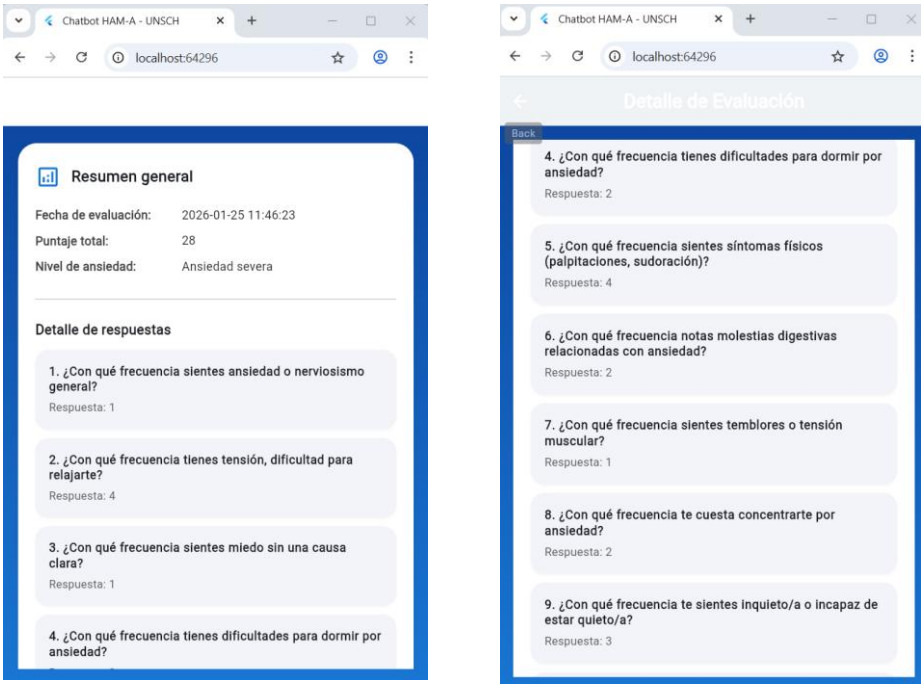
(a) (b)

Figure 14: Graphical user interface of the HAM-a generative AI chatbot: (a) question on irritability associated with anxiety and (b) results screen displaying anxiety severity classification and automated feedback



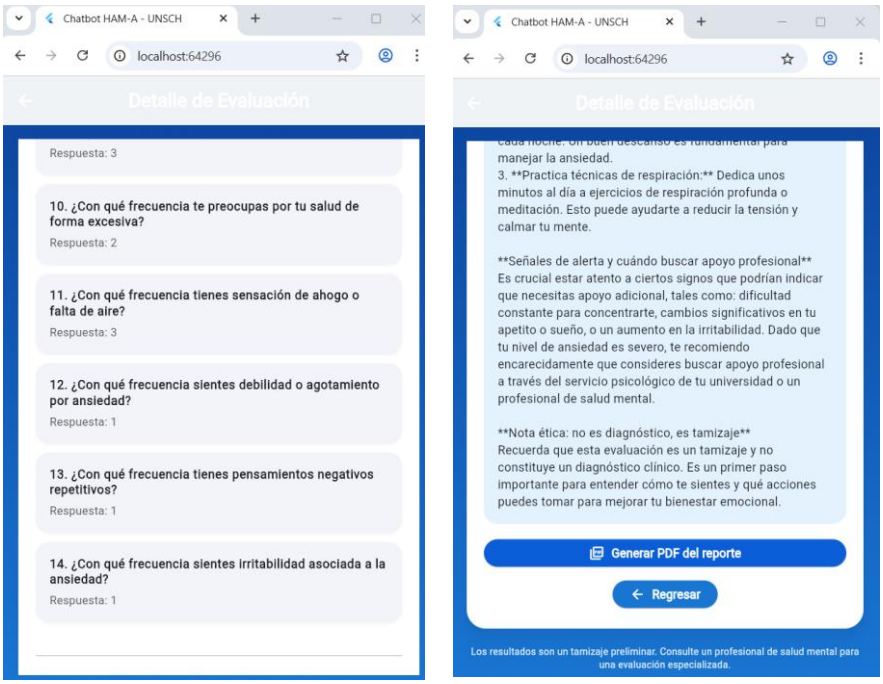
(a) (b)

Figure 15: Graphical user interface of the HAM-a generative AI chatbot: (a) results screen displaying total score and anxiety severity classification and (b) detailed screen presenting personalized AI-generated feedback and practical recommendations



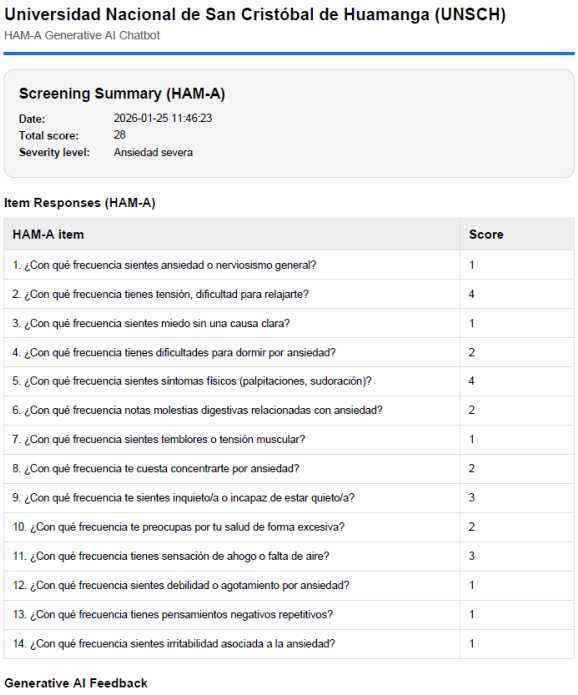
(a) (b)

Figure 16: Graphical user interface of the HAM-a generative AI chatbot: (a) evaluation summary screen showing date, total score, and anxiety severity level and (b) detailed screen displaying item-level responses from the HAM-a assessment



(a) (b)

Figure 17: Graphical user interface of the HAM-a generative AI chatbot: (a) detailed screen displaying remaining item-level responses of the HAM-a assessment and (b) detailed screen presenting full AI-generated feedback with options to generate the PDF report and return to the previous screen



Universidad Nacional de San Cristóbal de Huamanga (UNSCH)

HAM-A Generative AI Chatbot

Screening Summary (HAM-A)

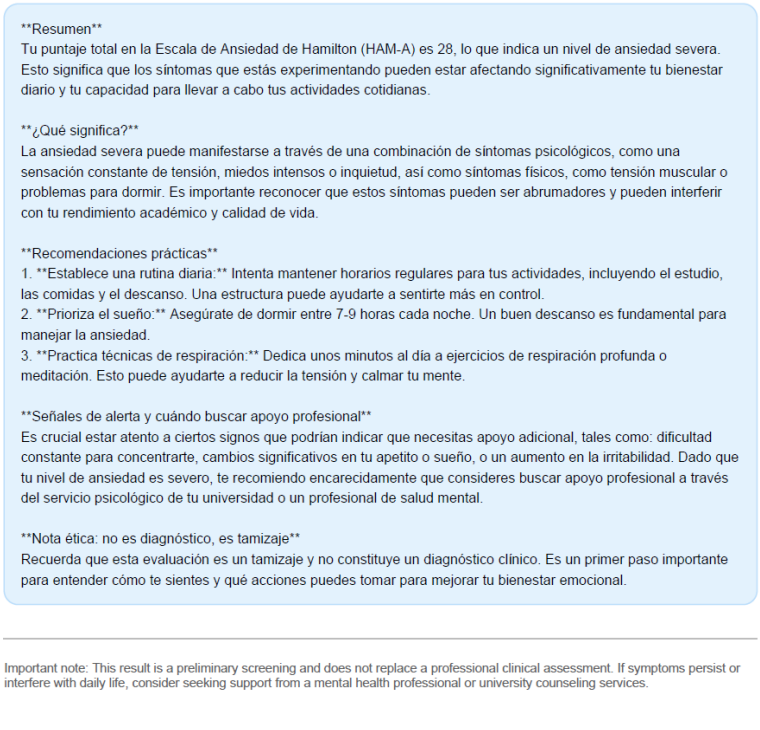
Date: 2026-01-25 11:46:23
Total score: 28
Severity level: Ansiedad severa

Item Responses (HAM-A)

HAM-A item	Score
1. ¿Con qué frecuencia sientes ansiedad o nerviosismo general?	1
2. ¿Con qué frecuencia tienes tensión, dificultad para relajarte?	4
3. ¿Con qué frecuencia sientes miedo sin una causa clara?	1
4. ¿Con qué frecuencia tienes dificultades para dormir por ansiedad?	2
5. ¿Con qué frecuencia sientes síntomas físicos (palpitaciones, sudoración)?	4
6. ¿Con qué frecuencia notas molestias digestivas relacionadas con ansiedad?	2
7. ¿Con qué frecuencia sientes temblores o tensión muscular?	1
8. ¿Con qué frecuencia te cuesta concentrarte por ansiedad?	2
9. ¿Con qué frecuencia te sientes inquieto/a o incapaz de estar quieto/a?	3
10. ¿Con qué frecuencia te preocupas por tu salud de forma excesiva?	2
11. ¿Con qué frecuencia tienes sensación de ahogo o falta de aire?	3
12. ¿Con qué frecuencia sientes debilidad o agotamiento por ansiedad?	1
13. ¿Con qué frecuencia tienes pensamientos negativos repetitivos?	1
14. ¿Con qué frecuencia sientes irritabilidad asociada a la ansiedad?	1

Generative AI Feedback

Figure 18: Automatically generated PDF report of the HAM-a anxiety screening: summary, item-level responses, and generative AI feedback for unsch students



Resumen
Tu puntaje total en la Escala de Ansiedad de Hamilton (HAM-A) es 28, lo que indica un nivel de ansiedad severa. Esto significa que los síntomas que estás experimentando pueden estar afectando significativamente tu bienestar diario y tu capacidad para llevar a cabo tus actividades cotidianas.

¿Qué significa?
La ansiedad severa puede manifestarse a través de una combinación de síntomas psicológicos, como una sensación constante de tensión, miedos intensos o inquietud, así como síntomas físicos, como tensión muscular o problemas para dormir. Es importante reconocer que estos síntomas pueden ser abrumadores y pueden interferir con tu rendimiento académico y calidad de vida.

Recomendaciones prácticas
1. **Establece una rutina diaria:** Intenta mantener horarios regulares para tus actividades, incluyendo el estudio, las comidas y el descanso. Una estructura puede ayudarte a sentirte más en control.
2. **Prioriza el sueño:** Asegúrate de dormir entre 7-9 horas cada noche. Un buen descanso es fundamental para manejar la ansiedad.
3. **Practica técnicas de respiración:** Dedica unos minutos al día a ejercicios de respiración profunda o meditación. Esto puede ayudarte a reducir la tensión y calmar tu mente.

Señales de alerta y cuándo buscar apoyo profesional
Es crucial estar atento a ciertos signos que podrían indicar que necesitas apoyo adicional, tales como: dificultad constante para concentrarte, cambios significativos en tu apetito o sueño, o un aumento en la irritabilidad. Dado que tu nivel de ansiedad es severo, te recomiendo encarecidamente que consideres buscar apoyo profesional a través del servicio psicológico de tu universidad o un profesional de salud mental.

Nota ética: no es diagnóstico, es tamizaje
Recuerda que esta evaluación es un tamizaje y no constituye un diagnóstico clínico. Es un primer paso importante para entender cómo te sientes y qué acciones puedes tomar para mejorar tu bienestar emocional.

Important note: This result is a preliminary screening and does not replace a professional clinical assessment. If symptoms persist or interfere with daily life, consider seeking support from a mental health professional or university counseling services.

Figure 19: Generative AI–based personalized feedback for HAM-a anxiety screening, including summary interpretation, practical recommendations, and ethical disclaime

4.2 Data preparation for agreement analysis using Cohen's kappa coefficient (κ)

Table 2 presents the distribution of the 50 Engineering students from the Universidad Nacional de San Cristóbal de Huamanga, Peru, enrolled in the 2025-II academic semester, who participated in the study. Cluster sampling was used to determine the sample, considering as grouping units the students belonging to specific academic groups or sections of the Engineering program. This sampling method was considered appropriate because the student population was naturally organized into academic clusters, which allowed the sample to be selected in a more accessible, orderly, and feasible manner for the application of the instrument. The data collection was conducted during the 2025-II academic semester as part of a completed pilot evaluation.

Each student completed the Hamilton Anxiety Rating Scale (HAM-A) through the generative artificial intelligence chatbot. In parallel, the HAM-A results were reviewed by two clinical mental health experts with experience in psychological assessment and the evaluation of anxiety symptoms. The experts established the reference clinical classification by consensus, assigning each case to one of three anxiety severity categories: mild, moderate, or severe. Table 1 presents the students' total HAM-A scores, the severity classification generated by the chatbot, and the consensus-based clinical classification assigned by the experts.

Table 2: Sample of 50 university students for anxiety severity classification using the HAM-a scale: comparison between generative AI chatbot screening and mental health expert assessment

ID	Total_Score	Chatbot_Class	Expert_Class
S01	35	Severe	Severe
S02	29	Severe	Severe
S03	30	Severe	Severe
S04	28	Severe	Severe
S05	24	Moderate	Moderate
S06	38	Severe	Severe
S07	30	Severe	Severe
S08	36	Severe	Severe
S09	31	Severe	Severe
S10	28	Severe	Severe
S11	22	Moderate	Moderate
S12	25	Severe	Mild
S13	34	Severe	Severe
S14	33	Severe	Severe
S15	32	Severe	Severe
S16	25	Severe	Severe
S17	28	Severe	Severe
S18	24	Moderate	Moderate
S19	30	Severe	Severe
S20	16	Mild	Mild
S21	36	Moderate	Moderate
S22	24	Moderate	Moderate
S23	22	Moderate	Moderate
S24	20	Moderate	Moderate
S25	25	Severe	Severe
S26	30	Severe	Severe
S27	30	Severe	Severe
S28	36	Severe	Severe
S29	24	Moderate	Mild
S30	31	Severe	Severe
S31	26	Severe	Severe
S32	34	Severe	Mild
S33	27	Severe	Severe
S34	27	Severe	Severe
S35	30	Severe	Severe
S36	21	Moderate	Moderate
S37	30	Severe	Severe
S38	26	Severe	Severe
S39	30	Severe	Severe
S40	20	Moderate	Moderate
S41	37	Severe	Severe
S42	26	Severe	Severe
S43	36	Severe	Moderate
S44	31	Severe	Severe
S45	28	Severe	Severe
S46	29	Severe	Severe
S47	24	Moderate	Moderate
S48	34	Severe	Severe
S49	30	Severe	Severe
S50	35	Severe	Severe

This information constitutes empirical judgment and comparative agreement analysis, enabling systematic comparison between the decisions of the automated system and professional clinical judgment, and supporting the calculation of Cohen's Kappa coefficient and the construction of the confusion matrix.

For each student, two categorical labels were recorded:

- **Expert_class:** severity level assigned by the clinical expert (mild, moderate, severe)
- **Chatbot_class:** severity level assigned by the chatbot (mild, moderate, severe)

Both labels represent a nominal classification with three mutually exclusive categories. The evaluation was framed as an inter-rater agreement problem between the "Expert vs. Chatbot," as well as a multiclass classification performance problem assessed through confusion matrix analysis and derived performance metrics.

4.3 Confusion matrix

The confusion matrix was constructed by cross-tabulating:

- Rows = true class (expert)
- Columns = predicted class (chatbot)

This approach allows for quantification of agreements (diagonal elements) and discrepancies (off-diagonal elements).

Table 3: Table confusion matrix (expert vs. Chatbot) for HAM-a severity classification (n = 50)

Expert \ Chatbot	Mild	Moderate	Severe	Total (Expert)
Mild	1	1	2	4
Moderate	0	10	1	11
Severe	0	0	35	35
Total (Chatbot)	1	11	38	50

As shown in Table 3, the main diagonal (1 + 10 + 35 = 46) represents the exact classification matches, indicating

a high level of concordance between the clinical expert and the chatbot. Misclassifications are predominantly concentrated in the Mild category, where the chatbot shows confusion with the Moderate (1 case) and Severe (2 cases) categories. In contrast, the Severe category exhibits no errors toward lower severity classes, suggesting a conservative system behavior: when the expert identifies high severity, the chatbot consistently preserves that classification.

4.4 Calculation of Cohen's kappa coefficient (κ)

Cohen's Kappa coefficient was employed to estimate the agreement between the chatbot and the clinical expert, allowing correction for the level of agreement expected by chance, according to formulas (1), (2), and (3). Let N denote the total number of cases and let CM represent the confusion matrix.

a. Observed agreement (P_0)

$$P_0 = \frac{\sum_{i=1}^{k} CM_{ii}}{N} \quad (1)$$

That is, the sum of the main diagonal divided by N.

b. Expected agreement by chance (p_e)

where R_i represents the total of row i (proportion per class according to the Expert) and C_i represents the total of column i (proportion per class according to the Chatbot).

$$P_e = \sum_{i=1}^{k} \frac{R_i C_i}{N^2} \quad (2)$$

Cohen's Kappa coefficient (κ)

$$\kappa = \frac{P_0 - P_e}{1 - P_e} \quad (3)$$

Substituting the values yields $P_0 = 0.9200$, $P_e = 0.5820$, and $\kappa = 0.8086$.

The agreement analysis between the anxiety severity classification generated by the generative artificial intelligence chatbot and the evaluation conducted by the clinical expert demonstrated a high level of consistency between both methods. The observed agreement reached $P_0 = 0.9200$, indicating that in 92.0% of the evaluated cases, both classifications exactly matched in assigning the level of anxiety. This result reflects strong direct correspondence and suggests that the chatbot closely reproduces the criteria applied in conventional clinical assessment.

However, since part of this agreement may be attributable to the distribution of categories rather than true concordance, the expected agreement by chance was estimated, yielding $P_e = 0.5820$. This coefficient represents the proportion of matches that could occur purely by chance, considering the relative frequency of mild, moderate, and severe anxiety categories within the analyzed sample. The obtained value highlights a non-uniform distribution of classifications, particularly influenced by the higher prevalence of certain severity levels. After adjusting the observed agreement for chance using Cohen's Kappa coefficient, a value of $\kappa = 0.8086$ was obtained. According to the criteria established by Landis and Koch (1977), who classify κ values above 0.80 as indicative of excellent agreement, this result reflects a near-perfect level of concordance. This finding suggests that the generative artificial intelligence–based system can identify and classifying anxiety severity with a high degree of consistency relative to expert clinical assessment.

4.5 Metrics derived from the confusion matrix

In addition to Cohen's Kappa coefficient (κ), the chatbot's performance was evaluated using multiclass classification metrics computed under the one-vs-rest approach, a fundamental strategy for analyzing models operating with more than two categories. In this framework, each class is evaluated independently by considering it as the positive class (one), while all remaining categories are grouped and treated collectively as the negative class (rest). Consequently, the multiclass problem is iteratively transformed into a series of binary classification problems, one for each anxiety severity level.

For each class, the following definitions apply:

- TP (true positives): cases of the class correctly predicted as such.
- FP (false positives): cases predicted as that class but belonging to another category.
- FN (false negatives): cases belonging to that class but predicted as a different category by the chatbot.
- TN (true negatives): cases not belonging to that class and not predicted as that class.

Table 4: Class-specific metrics (one-vs-rest approach)

Class	**TP**	**FP**	**FN**	**TN**
Mild	1	0	3	46
Moderate	10	1	1	38
Severe	35	3	0	12
Class	**Precision**	**Recall**	**Specificity**	**F1**
Mild	1.0000	0.2500	1.0000	0.4000
Moderate	0.9091	0.9091	0.9744	0.9091
Severe	0.9211	1.0000	0.8000	0.9589

Table 4 presents the class-specific analysis of severity levels, allowing the identification of distinct performance patterns of the chatbot. In the mild category, a precision value of 1.00 indicates that when the chatbot assigns this classification, it virtually does not commit errors, that is, it does not label students as mild if they are not classified as such by the clinical expert. However, the reduced recall (0.25) reveals that the system correctly identifies only one out of four actual mild cases, reclassifying the remaining three cases into higher severity levels, such as moderate or severe. This behavior is characteristic of models designed with a preventive orientation, prioritizing sensitivity

toward higher-risk levels and avoiding underestimation of severity, even at the cost of reduced detection of mild cases.

In the moderate category, precision and recall values close to 0.91 reflect balanced and consistent performance, with a limited number of errors both in overclassification (incorrectly labeling non-moderate cases as moderate) and under classification (failing to identify some true moderate cases). Additionally, the high specificity (0.9744) indicates that the chatbot rarely assigns this category when it is not warranted, suggesting adequate discrimination of this intermediate severity level.

In the severe category, a recall of 1.00 confirms that all cases classified as severe by the clinical expert were correctly identified by the chatbot, with no false negatives. This finding is relevant in screening and decision-support contexts, as it substantially reduces the risk of overlooking potentially high-priority cases. However, the precision (0.9211) and specificity (0.8000) values indicate the presence of some false positives—namely, non-severe students classified as severe by the chatbot. Although this may increase referral rates, in first-level screening systems it is generally preferable to prioritize sensitivity over strict specificity, because this approach supports the timely prioritization of students who may require professional evaluation.

Table 5 presents the overall performance metrics, providing a more precise understanding of the chatbot's behavior in a multiclass classification scenario. The accuracy (Accuracy = 0.92) indicates that the chatbot correctly classified 92% of cases, reflecting strong overall performance. However, this value may be influenced by class imbalance, as the severe anxiety category accounts for the largest proportion of cases. The Weighted F1-score (0.9032), which weights performance according to the support of each class, remains high primarily because the severe class (35 cases) dominates the average. While this metric effectively reflects real-world performance within the sample, it may obscure weaknesses in minority classes. In contrast, the Macro F1-score (0.756), which assigns equal weight to each class regardless of frequency, decreases mainly due to the low F1-score observed in the mild category (0.40). Similarly, the Balanced Accuracy (0.7197), defined as the average recall across all classes, is reduced primarily by the low recall of the mild category (0.25).

Table 5: Overall system performance metrics (n = 50).

Metric	Value
Accuracy	0.9200
Macro_F1	0.7560
Weighted_F1	0.9032
Balanced_Accuracy	0.7197
N	50

5 Discussions

The findings of this pilot study indicate that the generative artificial intelligence chatbot based on the HAM-A scale achieved a high level of agreement with expert classifications, suggesting its preliminary potential as a screening and decision-support tool in university mental health contexts. The obtained Cohen's Kappa coefficient ($\kappa = 0.8086$) falls within the range of almost perfect agreement according to Landis and Koch, indicating that the concordance between the automated system and the clinical expert exceeded chance expectations. However, given the pilot nature of the study, the limited sample size, and the single-university context, these findings should be interpreted cautiously and understood as preliminary evidence of feasibility and agreement rather than definitive evidence of clinical validity or broad institutional applicability.

From a comparative perspective, prior studies on AI-based chatbots in mental health have primarily focused on therapeutic effectiveness, symptom reduction, feasibility, and user engagement, rather than agreement with expert-based classification. For example, the generative AI chatbot Wayhaven demonstrated significant reductions in depression ($\beta = -1.62$; $p < .001$) and anxiety ($\beta = -2.15$; $p < .001$), alongside improvements in well-being ($t_{40} = 2.90$; $p = .006$; $d = 0.45$), indicating its feasibility and acceptability among university students [13]. Similarly, the chatbot Tess showed statistically significant reductions in anxiety ($p = .045$ and $p = .02$ across groups) and depression ($p = .03$) in a randomized controlled trial, supporting the usefulness of AI-driven psychological interventions [23].

In real-world settings, the Wysa chatbot demonstrated improvements in depressive symptoms, with high-engagement users achieving greater reductions (mean improvement = 5.84 vs. 3.52; $p = .03$; effect size = 0.63), as well as positive user acceptance (67.7% positive feedback), reinforcing its practical applicability [19]. Additionally, previous work with conversational agents such as Woebot has reported short-term reductions in anxiety and depression symptoms using CBT-based automated interventions [24]. These studies suggest that AI chatbots may support mental health-related processes through scalable, personalized, and accessible solutions; however, their role should be understood as complementary and not as a replacement for professional care.

A relevant distinction of the present pilot study is that it does not evaluate therapeutic outcomes or clinical diagnosis. Instead, it examines the level of agreement between chatbot-generated HAM-A severity classifications and the classifications assigned by a clinical mental health expert. Therefore, the contribution of this study lies in exploring classification concordance and screening support, rather than establishing diagnostic accuracy in a strict clinical sense. This perspective remains important because relatively few studies assess how

automated systems align with expert judgment using formal agreement metrics such as Cohen's Kappa.

The confusion matrix analysis shows that agreement is mainly concentrated along the main diagonal, particularly within the severe anxiety category, where the chatbot classified all cases identified as severe by the expert in the same category. In screening contexts, this behavior is relevant because it reduces the possibility of overlooking potentially high-priority cases. The recall value of 1.00 for the severe category suggests high sensitivity for this class within the analyzed pilot sample.

Nevertheless, the chatbot showed a tendency to overestimate severity in some cases, classifying certain mild or moderate cases as severe. This behavior may be acceptable in first-level screening contexts, where prioritizing sensitivity can be useful to avoid missing students who may require further professional evaluation. However, this conservative tendency also indicates the need for caution, since false positives may increase unnecessary referrals or concern among users. Therefore, the chatbot's classifications should be interpreted as screening outputs that require professional confirmation.

The class-specific metrics further support this interpretation. While the severe category showed strong performance, the mild category presented lower recall (0.25), indicating difficulty in distinguishing lower-severity cases within this sample. This result suggests that future versions of the system should improve discrimination between mild and moderate anxiety levels, particularly through threshold refinement, expert review of classification rules, and additional validation with more balanced datasets.

At the global level, the overall accuracy of 92% suggests favorable performance in this pilot evaluation. However, this metric should be interpreted cautiously due to class imbalance, as the severe anxiety category represents the largest proportion of cases. In this context, the Macro F1-score (0.756) and Balanced Accuracy (0.7197) provide a more conservative and informative interpretation, revealing variability across severity levels and identifying the mild category as an area for improvement.

From a technological standpoint, the integration of generative AI enables not only automated HAM-A administration and severity classification but also the generation of personalized feedback in natural language. This feature aligns with contemporary digital mental health approaches that emphasize accessibility, user-centered interaction, and guided self-understanding. Nevertheless, the AI-generated feedback should be carefully reviewed and framed with ethical disclaimers, since it must not be interpreted as professional diagnosis or individualized clinical treatment.

Despite the encouraging results, several limitations must be acknowledged. First, the study used a pilot sample of 50 students from a single university, which limits the generalizability of the findings. Second, the distribution of severity categories was imbalanced, which may have influenced the overall performance metrics. Third, the evaluation focused on agreement with expert classification, not on longitudinal outcomes, clinical diagnosis, or treatment effectiveness. Fourth, the quality, safety, and interpretability of the AI-generated feedback require further qualitative and expert-based assessment.

Compared with prior studies summarized in Table 1, the present pilot study differs in its evaluation focus. Previous AI-based mental health chatbots have commonly assessed therapeutic outcomes, symptom reduction, feasibility, engagement, or user acceptability [8-13,19,23,24]. In contrast, this study evaluates the concordance between chatbot-generated HAM-A severity classifications and clinical expert judgment. This distinction is relevant because agreement-based validation provides information about the consistency of automated classification outputs when compared with human expert assessment.

The obtained Cohen's Kappa coefficient ($\kappa = 0.8086$) and overall accuracy of 0.9200 suggest favorable agreement in this pilot sample. However, these results should not be interpreted as definitive clinical validation, since the sample was limited to 50 students from a single university. Compared with previous chatbot studies that primarily report changes in PHQ-9, GAD-7, engagement, or acceptability outcomes, the contribution of this work lies in integrating a structured anxiety scale, automated scoring, generative feedback, and expert-based agreement analysis within a single system. Cohen's Kappa value ($\kappa = 0.8086$) indicates high agreement beyond chance between the chatbot and the clinical expert in this pilot sample. This suggests that the chatbot's HAM-A severity classifications were generally consistent with expert judgment. However, the result should be interpreted as preliminary agreement, not diagnostic equivalence, and moderate or severe cases should be reviewed by qualified professionals.

Differences between the present findings and previous studies may be explained by several factors, including the population assessed, the use of HAM-A as the reference instrument, the classification-based design of the chatbot, and the evaluation method based on inter-rater agreement rather than longitudinal symptom change. The strong recall observed in the severe anxiety category may be useful for first-level screening and prioritization, but the low recall in the mild category indicates the need to refine thresholds and validate the system with more balanced datasets.

Several potential sources of bias should be considered when interpreting the findings. Selection bias may have occurred because participants were recruited from specific academic groups within a single university, while response bias may have influenced the self-reported HAM-A answers provided through the chatbot interface. In addition, class imbalance, particularly the predominance of severe cases, may have affected the overall performance metrics. Although these limitations were partially mitigated through the use of standardized HAM-A items, automated scoring rules, and expert-based

comparison, future studies should include broader sampling strategies, more balanced datasets, and multiple clinical evaluators.

Overall, the proposed chatbot should be considered a preliminary and complementary screening support tool for university mental health contexts. Its high agreement with expert classification in this pilot sample suggests potential usefulness for first-level anxiety severity screening, personalized feedback, and prioritization of students who may require professional attention. However, broader validation is necessary before recommending institutional or clinical deployment. Future research should include larger, more diverse, and preferably multicenter samples, as well as balanced severity groups, qualitative user evaluation, expert review of feedback, and comparison with additional standardized instruments.

6 Limitations and future work

This study has limitations inherent to a pilot evaluation, mainly the small sample size, its single-university setting, class imbalance, and the absence of longitudinal follow-up. Although the clinical classification was reviewed by two experts, no formal inter-rater reliability coefficient was estimated. Therefore, the results should be interpreted as preliminary evidence. Future research should validate the chatbot in larger, more diverse, and multicenter samples, include more balanced severity groups, estimate confidence intervals, assess usability, and compare the system with other standardized scales such as GAD-7, PHQ-9, or DASS-21.

7 Conclusions

The conclusions derived from the quantitative analysis and the design of the generative artificial intelligence chatbot suggest that the developed solution may function as a preliminary and complementary tool for anxiety severity screening and classification among university students using the HAM-A scale. The obtained Cohen's Kappa coefficient ($\kappa = 0.8086$) indicates a high level of agreement between the classifications generated by the chatbot and those assigned by the clinical expert within this pilot sample. This finding suggests that the automated system can approximate expert-based classification beyond agreement attributable to chance; however, it should not be interpreted as evidence of formal clinical diagnosis or definitive clinical validation.

The confusion matrix analysis shows that the highest level of agreement was concentrated in the severe anxiety category, in which the chatbot classified all cases identified as severe by the expert in the same category. This behavior is reflected in a recall of 1.00 for this class, suggesting that, within the analyzed pilot sample, the system did not miss severe cases. In first-level screening contexts, this result is relevant because it may support the prioritization of students who require professional attention. Nevertheless, the presence of false positives indicates that chatbot-generated classifications should be reviewed and confirmed by qualified mental health professionals.

Class-specific metrics calculated using the one-vs-rest approach provide a more nuanced understanding of the chatbot's performance. While the moderate and severe classes showed favorable precision, recall, and F1-score values, the mild class presented reduced recall, indicating a tendency to classify some mild cases into higher severity levels. This limitation suggests that future versions of the system should improve discrimination between mild and moderate anxiety levels. Global metrics, including accuracy (0.92), Weighted F1-score (0.9032), and balanced accuracy (0.7197), provide supportive evidence of preliminary performance, although these results must be interpreted cautiously due to the pilot design, limited sample size, and class imbalance.

From a design perspective, the chatbot developed in Flutter and Dart and integrated with a generative language model demonstrates the technical feasibility of combining modern software architectures with artificial intelligence techniques to support interactive, accessible, and user-centered screening processes. The generation of personalized natural language feedback based on HAM-A results adds value by helping students understand their screening outcomes; however, this feedback should be framed as informational and supportive, not as clinical advice or psychological treatment.

Overall, the obtained results and system design support the conclusion that the proposed solution has preliminary potential as a complementary tool for anxiety severity screening, classification, and decision support in university mental health contexts. The system is not intended to replace clinical evaluation but to support first-level screening and the prioritization of students who may require professional assessment. Due to the pilot nature of the study and the limited sample of 50 students from a single university, future research should validate the chatbot in larger, more diverse, and preferably multicenter samples before broader institutional or clinical implementation.

Acknowledgement

The authors express their gratitude to the National University of San Cristóbal de Huamanga for its academic and institutional support during the development of this research.

AI usage declaration

ChatGPT 5.0 Thinking was used as a generative artificial intelligence tool to support language refinement, academic editing, and feedback generation within the chatbot. The scientific content, analysis, results, and interpretations were fully reviewed and validated by the authors.

Author contribution statement

All authors contributed to the conception, design, development, analysis, writing, and revision of the manuscript. All authors approved the final version for publication.

References

[1] Li W, Zhao Z, Chen D, Peng Y, Lu Z. Prevalence and associated factors of depression and anxiety symptoms among college students: a systematic review and meta-analysis. Journal of Child Psychology and Psychiatry and Allied Disciplines. 2022; 63:1222-1230. doi:10.1111/jcpp.13606.

[2] Ahmed I, Hazell CM, Edwards B, Glazebrook C, Davies EB. A systematic review and meta-analysis of studies exploring prevalence of non-specific anxiety in undergraduate university students. BMC Psychiatry. 2023; 23:240. doi:10.1186/s12888-023-04645-8.

[3] Tan GXD, Soh XC, Hartanto A, Goh AYH, Majeed NM. Prevalence of anxiety in college and university students: an umbrella review. Journal of Affective Disorders Reports. 2023; 14:100658. doi: 10.1016/j.jadr.2023.100658.

[4] Riboldi I, et al. Digital mental health interventions for anxiety and depressive symptoms in university students during the COVID-19 pandemic: a systematic review of randomized controlled trials. Revista de Psiquiatría y Salud Mental. 2023; 16:47-58.

[5] Palmer CE, et al. Combining artificial intelligence and human support in mental health: digital intervention with comparable effectiveness to human-delivered care. Journal of Medical Internet Research. 2025;27: e69351. doi:10.2196/69351.

[6] Hua Y, et al. Charting the evolution of artificial intelligence mental health chatbots from rule-based systems to large language models: a systematic review. World Psychiatry. 2025; 24:383-394. doi:10.1002/wps.21352.

[7] Maier W, Buller R, Philipp M, Heuser I. The Hamilton Anxiety Scale: reliability, validity and sensitivity to change in anxiety and depressive disorders. Journal of Affective Disorders. 1988; 14:61-68. doi:10.1016/0165-0327(88)90072-9.

[8] Manole A, Cârciumaru R, Brînzaș R, Manole F. Harnessing AI in anxiety management: a chatbot-based intervention for personalized mental health support. Information. 2024; 15:768. doi:10.3390/info15120768.

[9] Zhang Q, Zhang R, Xiong Y, Sui Y, Tong C, Lin FH. Generative AI mental health chatbots as therapeutic tools: systematic review and meta-analysis of their role in reducing mental health issues. Journal of Medical Internet Research. 2025;27:e78238. doi:10.2196/78238.

[10] Mayor E. Chatbots and mental health: a scoping review of reviews. Current Psychology. 2025; 44:13619-13640. doi:10.1007/s12144-025-08094-2.

[11] Heinz MV, et al. Randomized trial of a generative AI chatbot for mental health treatment. NEJM AI. 2025;2(4). doi:10.1056/AIoa2400802.

[12] Chen C, et al. Comparison of an AI chatbot with a nurse hotline in reducing anxiety and depression levels in the general population: pilot randomized controlled trial. JMIR Human Factors. 2025;12: e65785. doi:10.2196/65785.

[13] Reyes-Portillo JA, et al. Generative AI-powered mental wellness chatbot for college student mental wellness: open trial. JMIR Formative Research. 2025;9: e71923. doi:10.2196/71923.

[14] Head KR. Minds in crisis: how the AI revolution is impacting mental health. Journal of Mental Health and Clinical Psychology. 2025; 9:34-44.

[15] Saraç H, Yüzakı E, Aşçı FH. The potential of AI chatbots as diagnostic tools in mental health: assessment of exercise dependence symptoms. Journal of Technology in Behavioral Science. 2025. doi:10.1007/s41347-025-00567-2.

[16] Wang Y, Wang Y, Crace K, Zhang Y. Understanding attitudes and trust of generative AI chatbots for social anxiety support. In: Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '25). New York, NY: Association for Computing Machinery; 2025. doi:10.1145/3706598.3714286.

[17] Casu M, Triscari S, Battiato S, Guarnera L, Caponnetto P. AI chatbots for mental health: a scoping review of effectiveness, feasibility, and applications. Applied Sciences. 2024; 14:5889. doi:10.3390/app14135889.

[18] Linardon J, Cuijpers P, Carlbring P, Messer M, Fuller-Tyszkiewicz M. The efficacy of app-supported smartphone interventions for mental health problems: a meta-analysis of randomized controlled trials. World Psychiatry. 2019; 18:325-336. doi:10.1002/wps.20673.

[19] Inkster B, Sarda S, Subramanian V. An empathy-driven, conversational artificial intelligence agent (Wysa) for digital mental well-being: real-world data evaluation mixed-methods study. JMIR mHealth and uHealth. 2018;6: e12106. doi:10.2196/12106.

[20] Rong G, Mendez A, Bou Assi E, Zhao B, Sawan M. Artificial intelligence in healthcare: review and prediction case studies. Engineering. 2020; 6:291-301. doi: 10.1016/j.eng.2019.08.015.

[21] Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nature Medicine. 2019; 25:44-56. doi:10.1038/s41591-018-0300-7.

[22] Torous J, et al. The growing field of digital psychiatry: current evidence and the future of apps, social media, chatbots, and virtual reality. World Psychiatry. 2021; 20:318-335. doi:10.1002/wps.20883.

[23] Fulmer R, Joerin A, Gentile B, Lakerink L, Rauws M. Using psychological artificial intelligence (Tess) to relieve symptoms of depression and anxiety:

randomized controlled trial. JMIR Mental Health. 2018;5: e64. doi:10.2196/mental.9782.

[24] Fitzpatrick KK, Darcy A, Vierhile M. 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. 2017;4: e19. doi:10.2196/mental.7785.

[25] Warrens MJ, de Raadt A, Bosker RJ, Kiers HAL. Weighted Kappa for interobserver agreement and missing data. Machine Learning and Knowledge Extraction. 2025; 7:18. doi:10.3390/make7010018.

[26] Hamilton M. The assessment of anxiety states by rating. British Journal of Medical Psychology. 1959; 32:50-55.

[27] Kessler RC, et al. short screening scales to monitor population prevalences and trends in non-specific psychological distress. Psychological Medicine. 2002; 32:959-976. doi:10.1017/S0033291702006074.

[28] Miner AS, Milstein A, Hancock JT. Talking to machines about personal mental health problems. JAMA. 2017; 318:1217-1218. doi:10.1001/jama.2017.14151.

[29] McHugh ML. Interrater reliability: the Kappa statistic. Biochemia Medica. 2012; 22:276-282. doi:10.11613/bm.2012.031.

[30] Bass L, Clements P, Kazman R. Software Architecture in Practice. 3rd ed. Boston, MA: Addison-Wesley; 2012.

[31] Fowler M. Patterns of Enterprise Application Architecture. Boston, MA: Addison-Wesley; 2003.

[32] Nyakhar S, Wang H. Effectiveness of artificial intelligence chatbots on mental health and well-being in college students: a rapid systematic review. Frontiers in Psychiatry. 2025; 16:1621768. doi:10.3389/fpsyt.2025.1621768.