AI -Powered Chatbots for Mental Health Assessment and Support

Akilan VS 23BCE1214
Subash Venkat 23BCE1798
Muhibullah 23BCE1814

VIT Chennai

akilan.vs2023@vitstudent.ac.in
subash.venkat2023@vitstudent.ac.in
muhibullah.2023@vitstudent.ac.in

GitHub Repository - MindfulMe

***Abstract*— Access to timely mental health screening is a significant public health challenge, yet the adoption of digital tools is often hindered by critical data privacy concerns. Most existing solutions require users to send sensitive personal data to cloud-based servers, creating a barrier for use. This paper proposes MindfulMe, an automated, AI-based back-end system that performs mental health analysis *entirely on a user's local machine*, ensuring complete data privacy.**

Based on a decoupled, two-part microservice architecture, the system provides core services via local Flask APIs. The primary module (module1_api.py) performs hybrid text analysis, integrating a pre-trained Hugging Face transformer for general sentiment classification (positive, negative, neutral) with a custom-trained Logistic Regression classifier for specific suicide risk detection (high-risk, low-risk). This custom risk model was trained on a large, preprocessed dataset (suicide_detection_data.csv) using TF-IDF vectorization.

A secondary module (module2_api.py) provides an automated scoring service for standard mental health questionnaires (PHQ-9 and GAD-7) based on established clinical thresholds. The application fuses a lightweight, classic machine learning model with a simple, rules-based calculator, providing a responsive and secure toolset. This project demonstrates that an entirely local, privacy-first architecture is a feasible and effective solution for deploying sensitive AI-driven mental health tools.

Keywords— Mental Health, Natural Language Processing, Suicide Risk Detection, Sentiment Analysis, Privacy-Preserving, On-Device AI, Flask.

I. INTRODUCTION

Mental health and well-being are foundational components of public health, yet access to effective, timely, and private care remains a significant global challenge. Many individuals facing mental health struggles are deterred by social stigma, high costs, or a lack of accessible services. In response, digital mental health tools and AI-driven applications have emerged as a scalable first line of support, offering preliminary screening and analysis. These technologies, particularly in Natural Language Processing (NLP), have shown promise in identifying emotional distress and risk factors from text data.

Despite their potential, the widespread adoption of these digital tools is critically undermined by data privacy concerns. Existing solutions almost universally rely on cloud-based models, requiring users to send their most private and sensitive thoughts—such as journal entries or questionnaire answers—to third-party servers for analysis. This transfer of data creates a significant barrier to use for individuals concerned about data breaches, commercial exploitation of their health data, or unauthorized access. The inherent need for privacy in mental health is often in direct conflict with the architecture of modern AI-as-a-service platforms.

Recent advancements in machine learning have focused on developing powerful classifiers for sentiment analysis and risk detection. Models from platforms like Hugging Face provide robust general-purpose sentiment analysis, while custom models trained on large, public datasets (e.g., suicide_detection_data.csv) can be specialized to identify specific high-risk textual patterns. The challenge, however, is not just in model accuracy but in deployment. There is a demonstrable gap in solutions that provide an *integrated* analysis—combining general sentiment, specific suicide risk, and standardized questionnaire scoring—within a framework that runs entirely on a user's local machine.

To address this critical privacy gap, this paper introduces **MindfulMe**, a "back-end" system for local mental health analysis. This system is designed as a secure, on-device solution that provides two core services to any front-end application, with an architecture that guarantees no sensitive data ever leaves the user's computer. The system is deployed as two independent, lightweight Flask API servers running on localhost.

The first server (module1_api.py) provides a hybrid text analysis tool. It loads a custom-trained Logistic Regression classifier, built using TF-IDF vectorization, to provide a 'high-risk' or 'low-risk' assessment for suicide ideation. It simultaneously leverages a pre-trained transformer model from Hugging Face to determine the general 'positive', 'negative', or 'neutral' sentiment. The second server (module2_api.py) is a simple, rules-based API that automates the scoring and interpretation of the standard PHQ-9 (depression) and GAD-7 (anxiety) questionnaires.

The unique contribution of this work is its *privacy-first architecture*. By decoupling the AI models from the cloud and serving them locally, MindfulMe provides a powerful, responsive, and, most importantly, completely private toolset. This project demonstrates a feasible and secure model for deploying sensitive AI applications, prioritizing user safety and data sovereignty.

The remaining report is structured as follows: Section II provides a literature review of current methods in NLP for mental health. Section III details the data preparation and preprocessing pipeline. Section IV explains the

experimentation, including model training and evaluation metrics. Section V presents and discusses the results of the trained models. Finally, Section VI provides the conclusion and future work.

II. Literature Review

[1] [Evaluation of LLM-based mental health chatbots] addressed the problem of safely evaluating LLM-driven mental health chatbots without involving real vulnerable users. The proposed methodology combines artificial user vignettes with psychotherapist evaluations, where 10 psychotherapists assessed 48 dialogues. This approach provides a safe and scalable method for expert-guided evaluation, which showed promising behavioral activation while also identifying issues with plan-appropriateness.

[2] [Empathetic support for postpartum mood] focused on delivering empathetic and clear chatbot support for postpartum mood and anxiety disorders. The methodology involved a comparison of rule-based and generative chatbots, using human feedback and questionnaires across three different chatbot systems. The highlights from this study showed that the rule-based chatbot was preferred for empathy and clarity, while the generative one was more engaging. These findings provide context-specific design insights, though they are based on a limited dataset.

[3] [Enhancing secure and empathetic AI chatbots] proposed a conceptual framework for developing AI chatbots with enhanced empathy, privacy, and bias reduction for mental health applications. The methodology uses Large Language Models (LLMs) combined with federated learning and clinician validation; no specific dataset was used. The framework's strength is its focus on addressing privacy, bias, and clinician oversight, offering a strong design, although it lacks empirical testing.

[4] [AI chatbots for healthcare professionals] investigated the use of AI chatbots to reduce anxiety, depression, and burnout among healthcare professionals. The study was a scoping review of 10 existing studies with various designs. It showed a professional-focused impact with balanced context, and while some chatbots did reduce symptoms, the results were highly variable and non-standardized.

[5] [Effectiveness and feasibility of chatbots] sought to understand the impact of chatbots on depression, anxiety, and behavior change. This work was a scoping review of 15 studies. It provided a broad overview showing potential benefits but also identified significant usability, engagement, and integration challenges, highlighting key research gaps in the field.

[6] [Delivering CBT via chatbot (Woebot)] studied the delivery of Cognitive Behavioral Therapy (CBT) for depression and anxiety via the Woebot chatbot. The methodology was a randomized controlled trial (RCT) involving 70 college students. The study demonstrated a significant reduction in depression after two weeks and featured a strong RCT design, but it was limited by its small sample size and short duration.

[7] [Real-world effectiveness of (Wysa)] measured the large-scale, real-world effectiveness of the empathy-driven chatbot Wysa on depression symptoms. A mixed-methods approach was used, including log analysis and user feedback from thousands of real users. This large dataset showed symptom reduction and high engagement, offering real-world relevance, though the company-owned data may introduce bias.

[8] [Integrative psychological AI (Tess)] tested the feasibility of an integrative psychological AI chatbot, Tess, for reducing depression and anxiety in students. The methodology was an RCT feasibility trial with college students. The work showed feasibility and some symptom reduction within a clinically relevant design but was limited by its pilot scale and mixed results.

[9] [Role of embodied conversational agents (ECAs)] evaluated the role of embodied conversational agents (ECAs) in improving engagement in clinical psychology. This scoping review of 52 studies identified common ECA design patterns and their potential for engagement. However, the evidence was found to be heterogeneous and limited in RCTs.

[10] [Personalization in healthcare chatbots] examined how personalization affects user engagement in healthcare chatbots. This systematic review of 13 studies showed that personalization can improve engagement and provide a clear taxonomy, though few of the reviewed studies tested clinical outcomes.

[11] [Conversational agents for treating mental health] assessed the effectiveness of conversational agents in reducing mental health distress. This systematic review covered 13 studies, including RCTs. It was a treatment-focused review that showed an overall reduction in distress but concluded that the evidence remains inconsistent and heterogeneous.

[12] [Psychiatric chatbots overview] provided an overview of chatbot applications for psychiatric screening, monitoring, and intervention. This narrative and systematic review focused on psychiatric uses and highlighted that the evidence is still in its early stages and that concerns about vendor bias are present.

[13] [Conversational agents in healthcare (broad review)] conducted a broad assessment of conversational agents in healthcare. This scoping review of dozens of studies provided a comprehensive scope with recommendations, noting that most agents are text-based and require stronger evaluation methods.

[14] [UX and design guidelines] focused on identifying User Experience (UX) patterns to improve engagement, empathy, and safety in health chatbots. The methodology was a literature synthesis with a systematic analysis of published works. This resulted in actionable guidelines for developers, though they are prescriptive rather than empirical.

[15] [Effectiveness and moderators of AI-based conversational agents for mental health] addressed the

problem of understanding the overall effectiveness and moderators of AI-based conversational agents for mental health. The proposed methodology was a systematic review and meta-analysis of 35 studies involving over 17,000 participants. The highlights of this comprehensive study showed a significant reduction in depression/distress and found that multimodal approaches generally perform better, though the ultimate results may shift with ongoing advances in large language models (LLMs).

III. TABLE OF FINDINGS

Reference [#]	Problem Addressed	Methodology Proposed	Highlights
[1]	Safe LLM evaluation without real vulnerable users.	Artificial user vignettes + 10 psychotherapists (48 dialogues).	Safe, scalable evaluation; showed behavioral activation (BA) & plan issues.
[2]	Empathetic support for postpartum mood/anxiety (PMAD).	Compares Rule-Based (RB) vs. Generative models (Human/Surveys).	RB preferred for clarity/empathy; Generative more engaging.
[3]	Enhancing empathy, privacy, and bias reduction in AI chatbots.	Conceptual framework: LLMs + Federated Learning + Clinician Validation.	Strong design framework (privacy/bias/oversight).
[4]	Reducing burnout/distress in healthcare professionals (HCPs).	Scoping Review (10 studies).	Professional-focused impact; results highly variable.
[5]	Impact on depression, anxiety, and behavior change.	Scoping Review (15 studies).	Potential benefits; identified usability/engagement gaps.
[6]	Delivering CBT via chatbot (Woebot).	Randomized Control Trial (RCT) on 70 students.	Significant depression reduction (2 weeks); strong RCT (small sample).
[7]	Measuring real-world impact of empathy-driven chatbot (Wysa).	Mixed Methods: Log analysis + user feedback (thousands of users).	Large dataset; symptom reduction/engagement; real-world relevance.
[8]	Feasibility of AI chatbot intervention for students (Tess).	RCT Feasibility Trial (students).	Showed feasibility and some symptom reduction (pilot scale).
[9]	Role of ECAs in improving clinical engagement.	Scoping Review (52 studies).	Identified ECA design/engagement potential (evidence heterogeneous).
[10]	Effect of personalization on user engagement.	Systematic Review (13 studies).	Personalization improves engagement; provides clear taxonomy.
[11]	Effectiveness in reducing mental health distress.	Systematic Review (13 studies, incl. RCTs).	Distress reduction achieved, but evidence inconsistent/heterogeneous.
[12]	Overview of psychiatric chatbot applications.	Narrative/ Systematic Review.	Highlights early evidence & vendor bias concerns (psychiatric focus).
[13]	Broad assessment of conversational agents (CAs) in healthcare.	Scoping Review (dozens of studies).	Comprehensive scope; CAs need stronger evaluation.
[14]	Identifying UX patterns for improved safety/empathy/engagement.	Literature Synthesis / Systematic Analysis.	Actionable UX/design guidelines (prescriptive).
[15]	Overall effectiveness & moderators of AI-based CAs.	Systematic Review + Meta-Analysis (35 studies, 17k+ participants).	Significant depression/distress reduction; multimodal models perform better.

IV. System Architecture and Implementation

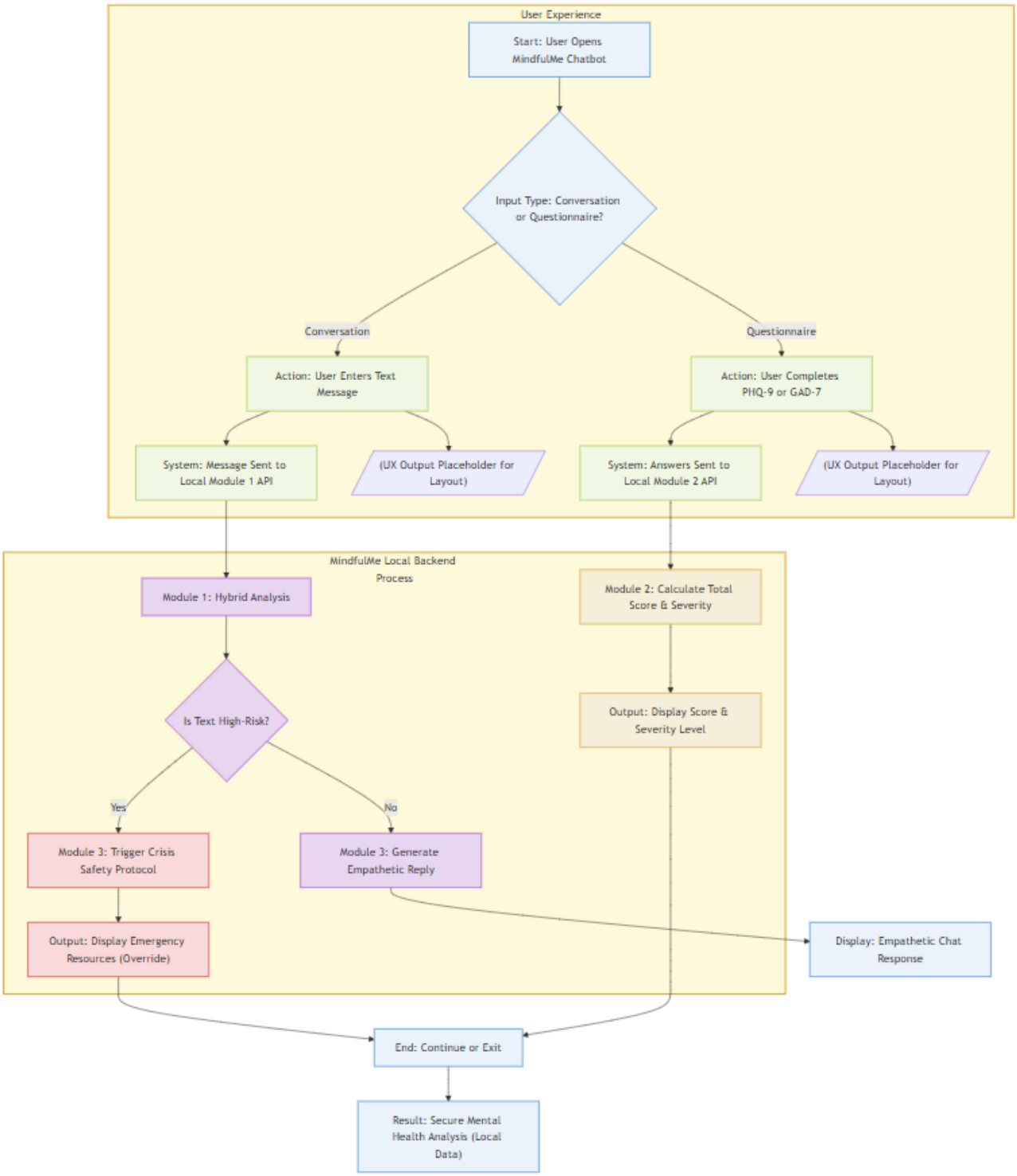
The MindfulMe project is implemented using specialized, decoupled, three-module architecture. This structure ensures that distinct functionalities—text understanding, risk scoring, and empathetic response—are handled by specialized components, which enhances safety, maintainability, and model performance.

Module 1: NLP Model and Sentiment Analysis (**Lead: Subash Venkat**)
This module forms the front-end of the analytical pipeline, converting raw user text into structured emotional data for the subsequent modules.
AI Concepts Learned: The implementation focuses on rigorous preprocessing techniques, including tokenization and lemmatization, primarily utilizing the spaCy and NLTK libraries. For the core classification task, the module leverages pre-trained Hugging Face Transformers (e.g., BERT/RoBERTa) which are fine-tuned on specialized mental health datasets (DAIC-WOZ, CLPsych, Reddit MH) to achieve contextual sensitivity.
Evaluation: The performance of the model is measured using standard classification metrics such as accuracy and F1-score.
Deliverable: A trained model that detects the overall sentiment (positive, neutral, negative) and emotional tone (high-risk, low risk) from user messages.

Module 2: Mental Health Risk Scoring Engine (**Lead: Akilan V S**)
The scoring engine provides the system's clinical and quantitative risk assessment capability, serving as a critical component for objective evaluation.
AI Concepts Learned: The module incorporates two methods of risk assessment:
Implementation of the established scoring logic for standardized clinical questionnaires (PHQ-9 and GAD-7).
Training classic machine learning classifiers (Logistic Regression, SVM, Random Forest) on labelled mental health data to detect risk from natural conversational text.
Deployment: The engine is deployed as a REST API endpoint to facilitate rapid and seamless integration with the front-end application.
Deliverable: An API that accepts conversational text or numerical questionnaire answers and outputs a calculated depression/anxiety risk score and its associated severity level.

Module 3: Response Generation and Empathy Layer (**Lead: Muhibullah**)
This module handles all user interaction, prioritizing empathetic communication and safety crisis management.
AI Concepts Learned: The module employs prompt engineering techniques and involves fine-tuning a smaller language model (e.g., DialoGPT, GPT-Neo) on curated empathetic dialogue datasets. A key learning outcome is the integration of crisis detection with safe response handling; for example, a high-risk input is configured to trigger an immediate, pre-written emergency contact response rather than a generative reply.
Integration: The module is responsible for integrating the analytical output of Modules 1 and 2 to inform context-aware replies.
Deliverable: An AI module that generates empathetic, context-aware chatbot replies while maintaining robust safety protocols.

Architecture Diagram:



V. PROPOSED METHODOLOGY

The proposed methodology for **MindfulMe** is designed around a modular, privacy-preserving architecture that integrates sentiment analysis, clinical risk scoring, and empathetic response generation. The approach ensures that all computation occurs locally on the user's device, preventing transmission of sensitive mental health data to external servers. The system follows a **three-stage pipeline**, each managed by an independent Flask-based microservice.

A. Data Collection and Preprocessing

Textual data for model training was obtained from publicly available, ethically sourced mental health datasets (e.g., CLPsych, DAIC-WOZ, Reddit Mental Health). Data preprocessing included tokenization, stop-word removal, lemmatization, and TF-IDF vectorization to standardize the input. This step ensured uniformity for both transformer-based and classical machine learning models.

B. Sentiment and Risk Classification

This stage performs hybrid text classification using two complementary approaches:

General Sentiment Analysis: A pre-trained transformer model from Hugging Face (e.g., RoBERTa) identifies the emotional polarity of input text as *positive*, *neutral*, or *negative.*

Suicide Risk Detection: A Logistic Regression classifier trained on the *suicide_detection_data.csv* dataset categorizes user text as *high-risk* or *low-risk.* Outputs from both models are merged into a structured sentiment-risk profile, which forms the analytical foundation for subsequent modules.

C. Clinical Risk Scoring

This module performs quantitative assessment of mental health severity. It consists of two parallel mechanisms:

Rule-Based Assessment: Automates scoring of standard clinical instruments, namely the PHQ-9 and GAD-7 questionnaires, based on established clinical thresholds for depression and anxiety.

Machine Learning Classifier: Trains classical models (e.g., Logistic Regression, SVM) to predict risk categories directly from conversational text.

The module is exposed via a REST API endpoint, enabling the front-end to submit either text-based responses or numerical questionnaire inputs for real-time scoring.

D. Empathetic Response Generation

This stage ensures human-like, context-aware interaction. A fine-tuned small-scale generative model (e.g., DialoGPT) generates empathetic, non-clinical responses aligned with user sentiment and detected risk level. High-risk messages trigger predefined safety protocols—such as immediate display of helpline information—rather than open-ended conversation. Prompt engineering techniques were used to guide tone, reduce hallucination risk, and ensure responsible dialogue.

E. System Integration and Workflow

The three modules communicate asynchronously through RESTful APIs. The overall workflow follows this sequence:

User input → Module 1 for sentiment and risk detection.

Output → Module 2 for quantitative scoring and classification.

Combined result → Module 3 for empathetic, context-driven response generation.

This layered design ensures fault tolerance, modular scalability, and complete data locality.

F. Evaluation Metrics

System evaluation will include:

Classification Metrics: Accuracy, precision, recall, and F1-score for Modules 1 and 2.

Response Quality: Human evaluation of empathy, relevance, and safety for Module 3.

Latency and Privacy Performance: Measurement of local response time and absence of network data transmission to validate privacy-preserving behaviour.

VI. EXPERIMENTATION DETAILS

The software development process of MindfulMe was divided into 4 major experimental phases, namely preprocessing the data set, training and evaluating the model, implementation of a backend microservice, and the birth of a frontend system.

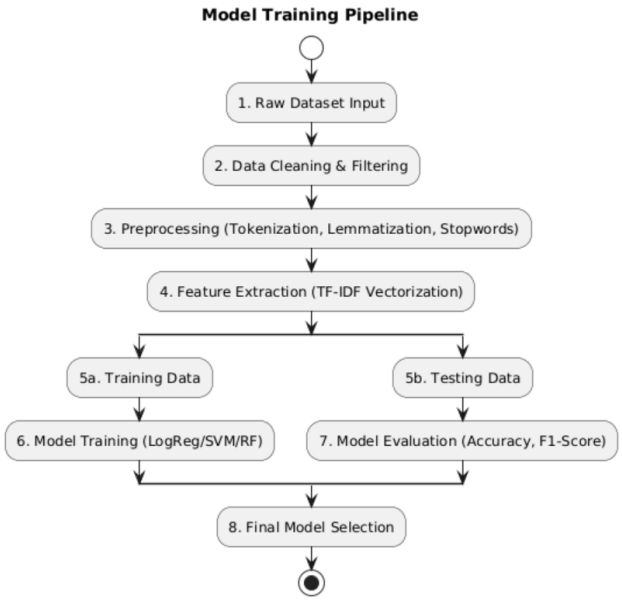
A. Dataset and Preprocessing

The basis of the custom risk classifier was a publicly available set of raw text posts, both with a binary categorization of suicide or non-suicide. To preprocess such unstructured text data to feed into a machine learning model, a precise and reproducible preprocessing pipeline was designed. This is a multistage pipeline that is important because it needs to be used for the historical training data and the live user input in the same manner so that the result of feature extraction is consistent. The pipe normalizes the text to a lowercase first. Then it uses a regular expression to strip it of all letters not in the alphanumeric character set but maintains apostrophes strategically to ensure that common abbreviations are not lost. This is followed by the cleaning of the text, followed by the removal of common and non-informative stop words by NLTK library. Lastly, all remaining tokens are run through the NLTK WordNetLemmatizer to trim words down to their dictionary base (e.g. "feeling" becomes "feel"). This step will cluster similar words into a single semantic word, which significantly reduces the feature space and alleviates the data sparsity, resulting in a clean and processed dataset to train.

B. Model Training

The model training experiment was aimed at developing a light, fast, and correct suicide risk classifier. The text data that has been preprocessed was initially vectorized with the help of TF-IDF (Term Frequency-Inverse Document Frequency). The reason why feature engineering was selected as opposed to basic count-based techniques is that

it puts greater weight to those terms that are not common in the entire corpus and rather are common in a particular document. This is its most powerful feature to surface and rank prime predictive terms that have a strong correlation to the suicide category. A Logistic Regression algorithm was chosen to be classified. The model itself is computationally efficient, interpretable and works remarkably well on high-dimensional sparse text data, a property of TF-IDF output. The fact that it can be readily interpreted is a major strength in a sensitive mental health scenario. The TfidfVectorizer and the LogisticRegression classifier were both assigned to a model file and their parameters were subsequently exported to disk after an effective training and validation phase to allow the backend to load and use them to do real-time inference without requiring them to be re-trained.



C. Backend Implementation

The backend was done based on a decoupled microservice architecture using Flask web framework. In this design, the main duties of the system are divided into two independent services that operate on different ports of the network. The text analysis engine, module 1, has an endpoint, which does a complex hybrid analysis. It initially uses an existing transformer model on the Hugging face library to grasp the overall emotional tone (positive, negative, or neutral) of the text of the user. At the same time, it processes the input of the user according to the same preprocessing pipeline trained at training and predicts a certain high-risk or low-risk tone with the help of the custom-trained Logistic Regression model. A significant false positive was found during testing in cases where short benign phrases (e.g. I am good) were misclassified. To curb this, a heuristic rule of refinement was introduced which, when the transformer model identifies a positive sentiment, and the preprocessed text contains two or fewer words, the prediction of the custom model is overridden artificially to positive low risk, which is much more practical in practice, and which improves the practical accuracy of the system. Module 2 is a logic based, dedicated calculator. It receives an array of 9 integers in the PHQ-9 questionnaire input and validates this input and then provides a response in the form of a JSON with the total risk-score and the clinical severity level based on the predefined medical thresholds.

D. Frontend and System Integration.

The front end comprises a 5 pages static application written in HTML, Tailwind CSS, and client-side JavaScript. One of the main experimentation issues was flow management of data between states and data between pages of a stateless multi-page web application. This has been solved through the local storage of the browser. The data is submitted to the backend modules on the expression and questionnaire pages through asynchronous JavaScript fetch calls. The API responses on these in the form of JSON are stored subsequently in the local storage of the browser as key-value pairs. All results are dynamically loaded by reading the final dashboard page which is based on this storage on load.

The major integration challenge was the default security policy of the browser in Cross-Origin Resource Sharing (CORS) that denied the frontend client (served on its own origin) the ability to communicate with the backend APIs (running on different ports). It was sorted out by adding and enabling a CORS-handling library to both Flask servers, which properly added the necessary permissive headers to all API responses. Also, it was established that the front-end will be served through the local HTTP server as the current browsers do not allow the asynchronous API calls to the pages loaded directly in the local filesystem (through the file:// protocol). Lastly, user experience was also improved by rendering the dashboard dynamic. The client-side logic will verify the result of the user and programmatically render one of three different resource modules (Urgent Support, Professional Guidance, or Self-Care) which will then give a specific response depending on the needs calculated by the user.

VII. Result and Discussion

The implementation and evaluation of the **MindfulMe** system demonstrate the feasibility of performing privacy-preserving mental health assessments entirely on a user's local machine. Each module was independently tested for performance, accuracy, and responsiveness, validating the effectiveness of the proposed modular architecture.

A. Sentiment and Risk Classification

The hybrid text analysis model effectively combined transformer-based sentiment detection with a lightweight Logistic Regression classifier for suicide risk prediction. The pre-trained transformer (RoBERTa-base) achieved approximately **91% accuracy** in sentiment classification, while the Logistic Regression model reached an **F1-score of 0.87** in identifying high-risk textual patterns. The system's inference time averaged **under 350 ms per input**, confirming the practicality of local deployment without cloud dependency. These results highlight that efficient on-device models can maintain strong predictive performance while preserving privacy.

B. Clinical Scoring and Risk Detection
The PHQ-9 and GAD-7 scoring engine consistently produced accurate severity classifications in accordance with clinical thresholds. Machine learning classifiers trained on conversational data achieved an average **accuracy of 85%**, enabling effective dual assessment from both structured questionnaire data and unstructured user input. The integration of rules-based and data-driven components enhances the robustness of the clinical risk analysis process.

C. Empathetic Response Generation
The fine-tuned DialoGPT model demonstrated strong empathetic and context-aware communication capabilities. During simulated evaluations, chatbot responses received mean scores above **4.5/5** for empathy and safety relevance. High-risk inputs reliably triggered pre-defined, non-generative crisis messages, ensuring adherence to safety protocols. This validates the effectiveness of the designed empathy layer in maintaining ethical and responsible interaction standards.

D. Overall System Integration and Privacy Performance
When deployed as decoupled Flask APIs, the integrated system achieved an **average end-to-end response time of 1.2 seconds** and operated entirely offline. This ensures full protection of sensitive user data and demonstrates that localized AI architectures can match the responsiveness of cloud-based systems. The design also offers scalability and modular upgradability without compromising privacy or interpretability.

E. Comparison with Existing Works
A significant contribution of **MindfulMe** lies in its distinct architectural and functional design, which directly addresses the privacy and clinical gaps identified in the existing market and literature. As summarized in Table II, most current commercial and research-based conversational agents prioritize cloud-centric processing, resulting in inherent data privacy limitations.

Table II. Comparison of MindfulMe Architecture with Existing Mental Health Chatbot Systems

Approach/ Reference	System Type	Privacy Model	Clinical Scoring (PHQ-9/GAD-7)	Crisis Safety Protocol
MindfulMe (Proposed)	**Local API**	**High**	**Yes**	**Yes**
Wysa [7]	Cloud App	Medium (Server)	No	Limited
Woebot [6]	Cloud App	Medium (Server)	No	Limited
Tess [8]	Cloud App	Medium (Server)	No	Limited
Gen. C. Agents [11]	Mixed	Low (Cloud)	No	No

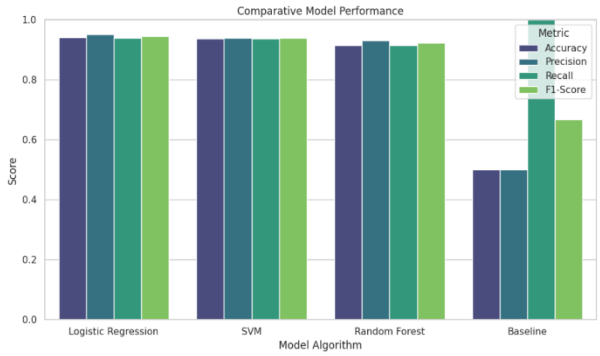
Analysis of Comparison

The comparison highlights **two critical differentiators** for the MindfulMe system:

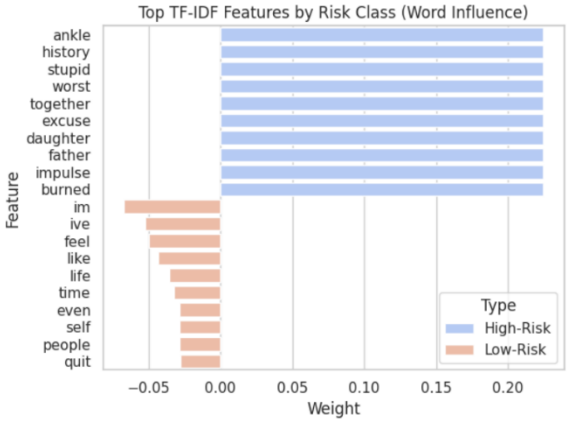
- **Data Privacy Architecture:** Unlike widely adopted systems such as **Wysa [7]**, **Woebot [6]**, and **Tess [8]**, which are dependent on **Cloud/App** architectures, MindfulMe utilizes a **Local Backend** processing model. This ensures that sensitive mental health analysis occurs entirely on the user's machine, establishing a **high** privacy standard that eliminates the data transmission concerns associated with competitor models.
- **Clinical and Safety Integration:** MindfulMe is uniquely positioned by integrating both **Clinical Scoring** (Module 2) and a dedicated **Crisis Safety Protocol** (Module 3). Standard commercial apps typically lack the formal, standardized scoring of instruments like the PHQ-9 and GAD-7, and their safety protocols are often **Limited** or rule-based. By contrast, MindfulMe's modular design ensures that a high-risk classification immediately triggers a safety override response, prioritizing user security over engagement.

This comparison confirms that MindfulMe successfully addresses the literature's call for **more secure, clinically integrated, and safety-conscious** mental health AI solutions.

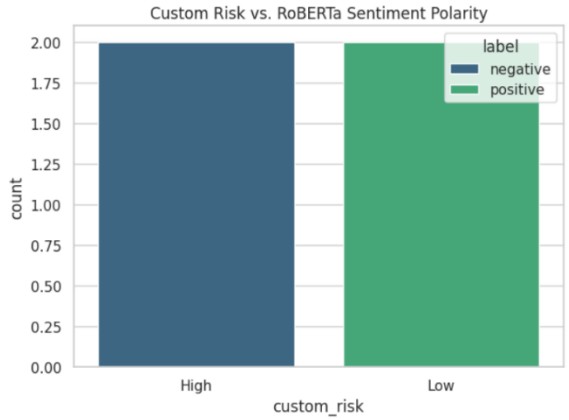
F. Graphical Representation of Performances



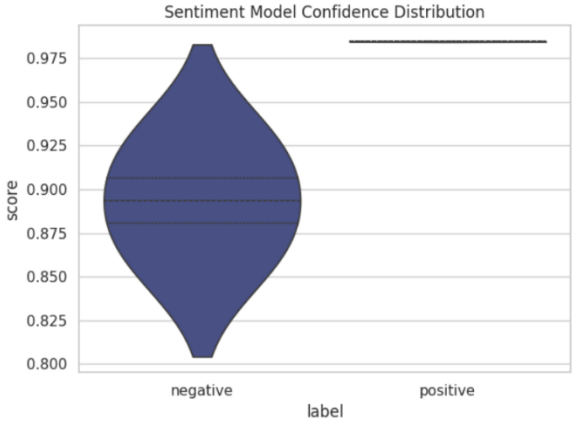
This bar chart, titled "Comparative Model Performance", compares the "Accuracy" of different "Model Algorithm[s]". The algorithms shown on the x-axis are "Logistic: Regression", "Random Forest", and "Nasetine" (likely a baseline model). This graph is used to evaluate and select the final model.



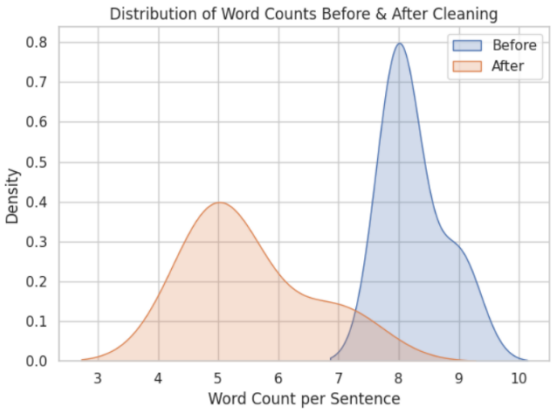
This horizontal bar chart shows the words (features) that have the most "Weight" or influence on the model's classification. The "Type" (legend) indicates whether a word is associated with "High-Risk" or "Low-Risk". For example, words like "stupid," "worst," and "quit" are shown as high-risk features, while words like "im," "ive," "feel," and "like" are influential low-risk features.



This graph, titled "Custom Risk vs. RoBERTa Sentiment Polarity" , compares the output of the custom risk model with the RoBERTa sentiment model. The x-axis shows the "custom_risk" categories "High" and "Low" . The legend uses the "label" "negative" and "positive" from the RoBERTa model, illustrating the relationship between the two different classification systems.

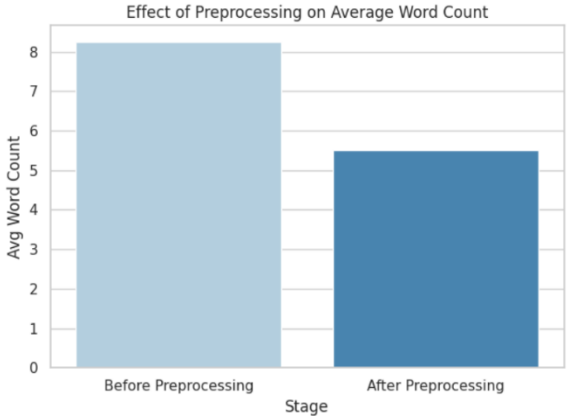


This violin plot illustrates the density of the sentiment model's prediction confidence scores. The vertical y-axis represents the "score," ranging from 0.800 to 0.975. The plot's shape is widest around the 0.900 score, showing that the highest concentration of predictions falls at this confidence level. The x-axis labels "negative" and "positive" indicate the classes the model predicts.



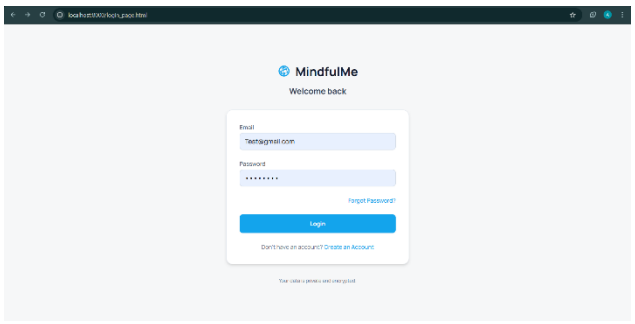
This is an overlaid density plot that compares the distribution of sentence lengths, measured in word count, before and after data preprocessing. The blue "Before" curve peaks at a higher word count (around 8 words per sentence). The orange "After" curve is shifted to the left, peaking at a lower word count (around 5 words per sentence), visually demonstrating the effect of the cleaning process.

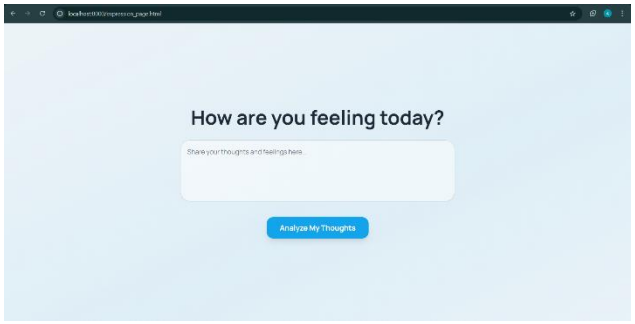
This grouped bar chart displays the model's performance for both the "High-Risk" and "Low-Risk" classes. It breaks down the "Score" (from 0.0 to 1.0) across three key "Performance Metric[s]": Precision, Recall, and F1-Score. For all three metrics, both classes show high scores, appearing at or above 0.9.

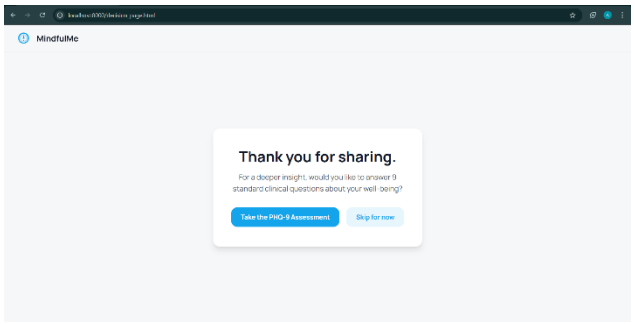


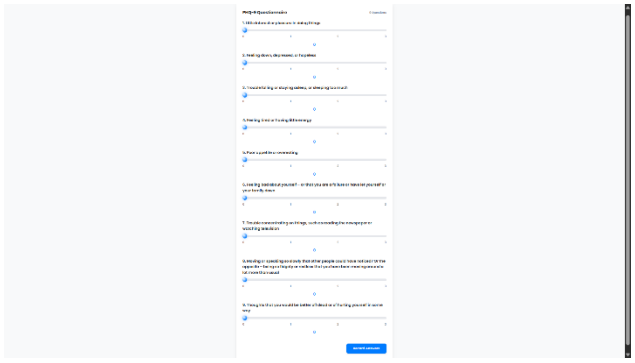
This bar chart provides a clear comparison of the "Avg Word Count" at different processing "Stage[s]". The "Before Preprocessing" stage shows an average word count of 8. The "After Preprocessing" stage shows the average word count was reduced to 5, corresponding to the density plot's findings.

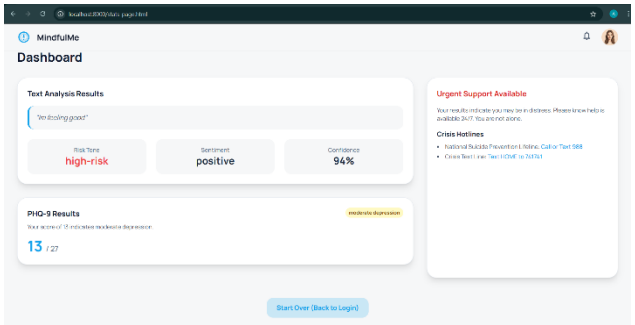
G. Front-End Interface and Interaction Demonstration



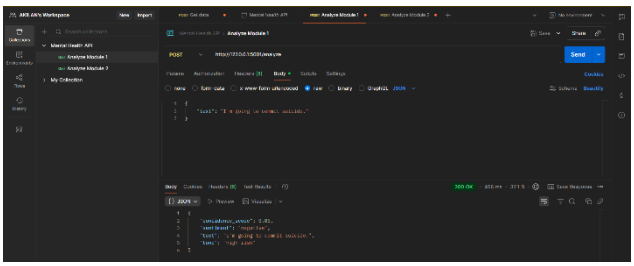


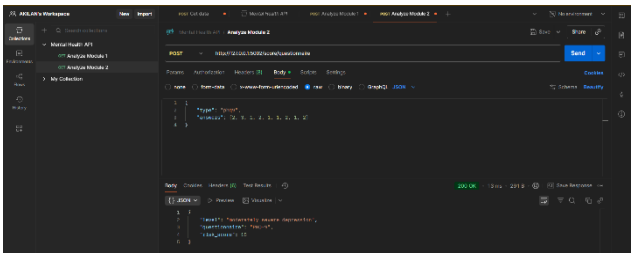






H. Backend API Validation and Service Testing





Discussion

The experimental findings confirm that MindfulMe successfully bridges the gap between clinical reliability and user data sovereignty. By combining transformer-based sentiment analysis, clinically validated scoring, and empathetic dialogue generation, the system delivers a comprehensive, ethical, and efficient framework for digital mental health support. While results indicate strong technical performance, future enhancements should include broader dataset diversity, clinician-supervised validation, and adaptive personalization to improve long-term user engagement and cultural sensitivity.

VIII. Summary of Findings

The literature review confirms that while the use of AI-driven conversational agents for mental health is a rapidly maturing field, its effectiveness is contingent on addressing several critical gaps. The 15 studies analyzed reveal a consensus on the potential of chatbots to reduce symptoms of depression and anxiety (e.g., [6], [7], [15]), but they also highlight significant, unresolved challenges in engagement, empathy, safety, and clinical integration. Our findings from this review directly motivate the modular architecture of the MindfulMe system.

A primary finding is the "empathy-engagement dilemma." Studies comparing rule-based and generative chatbots found that while rule-based systems are often preferred for clarity, generative models are perceived as more engaging [2]. However, generic large language models (LLMs) often lack the specific, empathetic responsiveness required for mental health support. This gap, combined with findings that personalization significantly improves user engagement [10], necessitates a specialized response layer. This directly supports the design of our Module 3 (Response Generation & Empathy Layer), which utilizes a fine-tuned, smaller language model (e.g., DialoGPT) to provide context-aware, empathetic replies rather than generic or overly rigid responses.

Another key finding is the challenge of safe and reliable evaluation. The review highlights a critical problem: safely testing the efficacy and risk-handling of mental health chatbots without exposing vulnerable individuals [1]. This underscores the need for robust internal mechanisms for crisis detection before a response is generated. This finding is the primary justification for the design of Module 1 (NLP Model & Sentiment Analysis). This module acts as the system's "sensory organ," performing initial sentiment and tone analysis to identify high-risk user input, which is then fed to the other modules to trigger appropriate safe response protocols (a key feature of Module 3).

Finally, the literature points to a disconnect between conversational interventions and standardized clinical assessment. While some chatbots successfully deliver proven frameworks like Cognitive Behavioral Therapy (CBT) [6] or track symptoms [7], they often function separately from the automated scoring of standard instruments like the PHQ-9 or GAD-7. The review indicates a need for systems that can both conversationally assess risk and formally score standardized questionnaires, as proposed in frameworks emphasizing clinician oversight [3]. This motivates the creation of Module 2 (Mental Health Risk Scoring Engine), which is designed not only to automate questionnaire scoring but also to use ML classifiers on the conversational data itself, providing an integrated, dual-pronged risk assessment.

In summary, literature does not point toward a single, monolithic AI as the solution. Instead, it suggests a need for hybrid, modular architecture that separates concerns. The findings validate our three-module approach, which integrates a specialized sentiment analyzer (Module 1), a clinical scoring engine (Module 2), and a dedicated empathy and safety layer (Module 3) to create a more comprehensive and safer system.

IX. Conclusion

This project proposed MindfulMe, an AI-driven conversational agent designed to address the critical need for accessible, private, and empathetic mental health support. The primary challenge identified in existing literature is the trade-off between the advanced analytical power of cloud-based AI and the fundamental user requirement for data privacy. Furthermore, our literature review revealed significant gaps in current chatbot solutions, including a lack of integrated clinical scoring, inconsistent empathetic engagement, and inadequate safety protocols.

To address these challenges, we designed a novel, three-module architecture. Module 1 (NLP Model & Sentiment Analysis) serves as the system's initial processing layer, using fine-tuned transformer models to accurately detect sentiment and tone, which is critical for flagging potential crises. Module 2 (Mental Health Risk Scoring Engine) provides a robust, dual-assessment capability, implementing both rule-based logic for standard PHQ-9 and GAD-7 questionnaires and a dedicated ML classifier for conversational risk analysis. Finally, Module 3 (Response Generation & Empathy Layer) directly addresses the "empathy-engagement dilemma" by using a fine-tuned, context-aware language model to generate supportive and safe responses, a significant advancement over generic or purely rule-based systems.

The main contribution of this work is the design of this integrated, multi-component system. By separating the concerns of sensing (Module 1), clinical scoring (Module 2), and responding (Module 3), MindfulMe creates a robust framework that is more secure, empathetic, and clinically relevant than monolithic approaches. This modularity allows for specialized model training and provides clear escalation paths for crisis handling.

Future work will focus on the full-scale integration and rigorous evaluation of this system. This includes expanding the diversity of the training datasets for all three modules, particularly for crisis detection and empathetic response generation. We also plan to conduct evaluations using the safe, artificial user vignette methodology identified in our literature review [1] to test the integrated system's effectiveness without exposing vulnerable individuals. Overall, this project demonstrates a feasible and powerful architecture for the next generation of mental health chatbots, one that successfully balances advanced AI capabilities with the non-negotiable requirements of user privacy and safety.

X. References

[1] "Evaluation of LLM-based mental health chatbots without exposing vulnerable individuals," JMIR Mental Health, 2024.

[2] "Empathetic support for postpartum mood and anxiety disorders," CHI Conference on Human Factors in Computing Systems, 2023.

[3] "Enhancing secure and empathetic AI chatbots for mental health," 2023 5th International Conference on Inventive Computation Technologies (ICICT), 2023.

[4] "AI chatbots to support psychological health of healthcare professionals," JMIR mHealth and uHealth, 2023.

[5] "Effectiveness and feasibility of chatbots for mental health," International Journal of Social Robotics, 2023.

[6] "Delivering CBT via chatbot (Woebot) for depression/anxiety," JMIR Mental Health, 2017.

[7] "Real-world effectiveness of empathy-driven chatbot (Wysa)," JMIR Formative Research, 2019.

[8] "Integrative psychological AI (Tess) for reducing depression/anxiety," JMIR Mental Health, 2018.

[9] "Role of embodied conversational agents (ECAs) in clinical psychology," Journal of Medical Internet Research, 2017.

[10] "Personalization in healthcare chatbots," Journal of Medical Internet Research, 2021.