SoulSync: An Intelligent Retrieval Interface for Emotional Support

Team 16

Sabhari Parameswaran
University of Illinois Urbana Champaign
sabhari2@illinois.edu

Mihir Shah
University of Illinois Urbana Champaign
mihirss3@illinois.edu

Omkar Dhekane
University of Illinois Urbana Champaign
dhekane2@illinois.edu

Harsh Premraj Vidhya
University of Illinois Urbana Champaign
harshpv2@illinois.edu

ABSTRACT

SoulSync is a web application that gives people private, on-demand emotional support using artificial intelligence. Users can type a message, speak into their microphone. The system then looks at facial expressions, voice tone, and the words the user shares to understand how they feel. Under the hood, SoulSync uses an open source vision language model called InternVL to turn these inputs into numerical “embeddings” which capture the user’s emotional state. These embeddings are compared against a Pinecone vector database of past conversations and mental health prompts to find the most relevant context. InternVL then generates a caring and helpful response, which is shown on screen and spoken aloud through ElevenLabs’ text-to-speech engine.

SoulSync runs entirely in real time and stores only anonymous session data, so each user can revisit earlier exchanges without sharing personal details. Our experiments show that the tool retrieves context with high precision and produces responses that users find empathetic and relevant. Typical response time is under two seconds, making the experience feel smooth and conversational. The source code is available on GitHub, and detailed instructions help anyone set up the system locally or in the cloud. By combining quick analysis, long-term memory, and natural voice feedback, SoulSync offers a simple yet powerful “virtual counselor” that users can turn to anytime they need support. Github: https://github.com/harshpv07/CS510

KEYWORDS

Information Retrieval, Large Language Models, AI agents, Mental Health, Multimodal.

INTRODUCTION

The global mental health crisis has reached unprecedented levels, with nearly one billion people suffering from mental disorders as traditional healthcare systems struggle with accessibility, stigma, and resource limitations. According to the World Health Organization, almost one in six adults will face depression at some point in their lives in the United States, and even more struggle with anxiety disorders. This care gap has accelerated innovation in artificial intelligence (AI), where systems now demonstrate remarkable capabilities in emotion recognition, contextual understanding, and therapeutic interaction. While existing solutions like text-based chatbots and mood tracking apps have shown promise, they often lack the multimodal perception and contextual awareness necessary for nuanced emotional support - limitations that become critical when handling complex human experiences like anxiety or depression.

Current AI mental health tools face three key challenges: (1) over-reliance on single-modality inputs (typically text) that miss critical emotional cues in voice and facial expressions, (2) static response systems lacking longitudinal context from previous interactions, and (3) privacy concerns that deter users from engaging fully. Research indicates that perceived humanness - characterized by consistent personality, memory of past exchanges, and multimodal responsiveness - significantly impacts users' acceptance of AI emotional support. The emerging field of vision-language models (VLMs) offers new opportunities to address these limitations through integrated analysis of verbal and non-verbal communication channels.

We propose SoulSync, an open-source emotional support system that combines real-time multimodal analysis with contextual memory retrieval. Our architecture advances current practices through three innovations: First, the integration of InternVL's vision-language capabilities enables simultaneous processing of text, vocal prosody, and facial microexpressions - a critical improvement over single-modality approaches. Second, our Pinecone vector database implementation creates dynamic user profiles through session-based embeddings, allowing the system to maintain therapeutic continuity without storing personal identifiers. Third, the sub-second response latency paired with ElevenLabs' expressive speech synthesis achieves conversational fluidity that users perceive as more authentic and engaging compared to batch-processing systems. We discuss our proposed methodology in subsequent sections.

MOTIVATION

Traditional therapy often comes with high costs and long waitlists. In rural areas, nearly half of Americans live without enough mental health professionals nearby, so it can be hard just to find a therapist. On top of that, many people worry about being judged or having their privacy compromised, which stops them from reaching out for help even when they need it most.

Some existing apps try to help by offering self help exercises or simple chatbots, but they miss a key part of human connection. They can read your words but not your tone of voice or your facial expressions. We created SoulSync to fill that gap. By listening to your voice, watching your face, and reading your words, SoulSync can sense how you really feel and give warm, human-like support right away. It stays available 24/7, keeps your data anonymous, and remembers past conversations so that each session feels personal and caring.

INTENDED USERS

SoulSync is designed for individuals seeking emotional support and self-assessment in a private and judgment-free environment. The tool is especially beneficial for :

- Everyday users who want to check in on their mental well-being and receive personalized feedback.
- Individuals facing barriers to traditional therapy, such as cost, stigma, or limited access to professionals .
- Counselors and researchers looking for a non-invasive screening tool to understand emotional patterns.
- People in remote or underserved communities who may not have consistent access to mental health services.

SoulSync serves as a helpful companion for anyone looking to better understand and manage their emotional health.

SYSTEM OVERVIEW

SoulSync is a web-based AI companion built to offer private, on-demand emotional support. You simply visit the SoulSync website with no login or personal details required and describe how you're feeling. The system runs around the clock and treats every session anonymously, so you can open up without worry. Whether you choose to type your thoughts, speak into your microphone, SoulSync accepts all of these inputs and listens attentively to your needs.

Once you share your message, SoulSync uses a powerful AI model called InternVL to analyze your input. InternVL is designed to understand text, speech, and images all at once. It reads your words, picks up on the tone of your voice, and even watches your facial expressions. By combining these signals with what you say, how you say it, and how you look, InternVL figures out your emotional state much like a human counselor would.

To make its support more personal, SoulSync keeps a private "memory" of past conversations in a high-speed storage called a Pinecone vector database. After each chat, your words and SoulSync's responses are converted into stored data points. When you return, the system searches this memory for moments that are similar to your current feelings, perhaps recalling when you last spoke about stress at work. This intelligent retrieval ensures that past context informs every new response, making the advice feel tailored and relevant.

Finally, after gathering your current input and the most relevant past memories, InternVL generates a thoughtful, caring reply in text form. SoulSync then shows you that message on-screen and simultaneously uses ElevenLabs' text-to-speech engine to read it out loud in a warm, natural voice. All of this happens within seconds, giving you a smooth, real time conversation experience. And throughout, your privacy remains protected, no names or identifiers are ever stored, so you can feel safe and supported every time you reach out.

CORE FEATURES

SoulSync provides several key features for emotional self-support:

- **Multimodal Input (text, audio, video):** SoulSync can take many forms of input. You can type a message, speak into the microphone. This means the app really listens to how you say things as well as what you say. By combining text, tone of voice, and facial cues, SoulSync gets a fuller picture of your feelings. Many apps only use text, but SoulSync goes beyond to understand you better through voice and video.

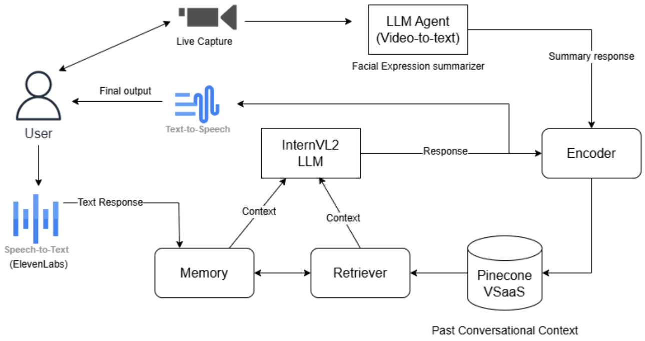
- **Emotion Detection:** SoulSync isn't just reading your words; it's also paying attention to your tone and expressions. It analyzes the sound of your voice and your facial expressions (and the sentiment of your words) to guess your mood. For example, it might notice if your voice sounds shaky or your face looks down, and use that to infer you're upset. This lets the system respond to how you truly feel, not just the literal text you typed, making its support feel more understanding.

- **Personalized Feedback:** Instead of giving stock answers, SoulSync offers support that's tailored to you. It uses an AI 'virtual therapist' to provide text (and even audio) feedback that fits the context of your conversation. This means it remembers details and adjusts its advice based on what you said earlier. Over time, SoulSync learns your story and needs, so its responses feel personal and caring much more than a one-size-fits-all reply.
- **Audio Output (real-time emotional text-to-speech):** SoulSync can speak to you in a warm, natural voice. We've integrated a text-to-speech engine (ElevenLabs) so that its replies aren't just shown on screen, but are read aloud with friendly inflection. The voice even carries gentle emotion, making the conversation feel more human. Hearing a comforting tone can make the experience more soothing than just reading plain text, which sets SoulSync apart from apps that are silent or monotonous.

- **Persistent Memory (session recall using Pinecone):** SoulSync doesn't forget your past chats. It stores important details (using a memory database called Pinecone) so that when you return later, it can recall what you talked about before. For instance, if you mentioned feeling anxious last week, SoulSync might ask how you're doing today. This ongoing memory makes interactions feel continuous and personal, unlike most apps that start fresh each time you open them.

- **Privacy and Accessibility (no sign-in, available anytime):** SoulSync is designed to be private and always available. You don't need to create an account or log in, so you can use it anonymously whenever you want. It's free to use 24/7, meaning support is just a click away any time of day or night. This anonymity and round-the-clock access help reduce stigma and make the service feel safe and welcoming unlike many apps that require sign-ups or have limited hours.

SYSTEM ARCHITECTURE



The system architecture integrates multimodal inputs and advanced language models to enable natural, context-aware conversational interactions between users and AI. The following components and data flows are central to the design:

1. *User Interaction Layer*:
   The user initiates interaction through spoken input, which is captured and processed by a Speech-to-Text service (ElevenLabs). This service converts the user's speech into text, enabling downstream processing.

2. *Live Video Capture and Expression Analysis*:
   Simultaneously, live video capture records the user's facial expressions. This video stream is processed by an LLM Agent (Video-to-text), which summarizes the user's facial expressions and emotional cues into a textual summary. This summary provides additional context to enhance the system's understanding of the user's intent and affective state.

3. *Core Language Model Processing:*
   The core of the system is the InternVL2 LLM, which receives both the transcribed user input and

the facial expression summary. The LLM synthesizes these multimodal inputs to generate a contextually relevant response. The LLM also interacts with two key context modules: 1) *Memory*: Stores recent conversational exchanges, enabling the system to maintain short-term context. 2) *Retriever*: Interfaces with both Memory and an external vector database (Pinecone VSaaS) to fetch relevant past conversational context, ensuring continuity and coherence in long-term interactions.

4. Context Management:
   The Pinecone VSaaS component serves as a scalable vector database, storing embeddings of past conversations. The Retriever queries this database to provide the LLM with pertinent historical context, which is essential for maintaining a personalized and context-aware dialogue.

5. Response Generation and Output:
   Once the InternVL2 LLM generates a response, it is passed to an Encoder module, which may further process or format the output as needed. The final response is converted from text back to speech using a Text-to-Speech module, delivering a natural, spoken reply to the user.

This architecture ensures robust, multimodal conversational AI that leverages both verbal and non-verbal cues, maintains conversational memory, and delivers seamless, human-like interactions.

EVALUATION

We evaluated our Retrieval-Augmented Generation (RAG) system across three core dimensions: Retrieval Efficiency, Generation Quality, and Response Time.

Retrieval Efficiency assesses the system's ability to accurately surface relevant prior conversations in response to a user query. This is measured using precision and recall, both before and after language model (LLM) augmentation.

Generation Quality reflects the final output produced by the language model, evaluated in terms of accuracy, relevance, completeness, usefulness, and fluency.

Response Time captures the end-to-end latency of the RAG pipeline. Latency is influenced by various factors, including network conditions, the number of tokens retrieved, and system architecture.

To quantify these metrics, we employed RAGAS, an open-source framework for evaluating RAG-based LLM systems. RAGAS provides detailed measurements across dimensions such as Faithfulness, Answer Relevance, Context Recall, and Context Precision. Upon running the benchmark using this framework, we obtained the following results:

Model	**Answer Relevance**	**Faithfulness**	**Context Precision**
OpenAI's GPT 4-0	0.800	0.5546	0.423
Google Gemini 2.5 Pro	0.872	0.834	0.542
Intern-VL 2.5B	0.816	0.754	0.742

Tab 1. Retrieval Efficiency

We evaluated our RAG pipeline, powered by Intern-VL 2.5B, against OpenAI's GPT-4.0 and Google's Gemini 2.5 Pro using the RAGAS evaluation framework. The results indicate that our system performs competitively across all key metrics. In terms of Answer Relevance, Gemini 2.5 Pro achieved the highest score (0.872), with Intern-VL closely following at 0.816, and GPT-4.0 slightly behind at 0.800. For Faithfulness, which measures factual consistency between the context and the generated response, our model scored 0.745, significantly higher than GPT-4.0 (0.5546) and approaching Gemini's 0.834. Notably, in Context Precision, which assesses the relevance and focus of the retrieved context, Intern-VL outperformed both commercial models with a score of 0.742, compared to Gemini's 0.542 and GPT-4.0's 0.423. These results demonstrate that our pipeline achieves a strong balance between generation quality and retrieval effectiveness, with particular strength in surfacing precise and relevant contextual information.

FUTURE WORK

One of the key limitations of current frontier LLMs lies in their inability to accurately perceive or respond to the emotional state of the user. Despite their linguistic capabilities, these models lack true emotional intelligence and contextual empathy, which are critical for building more human-centric interactions. Understanding and adapting to a user's emotional state remains a challenging task.

In future work, we aim to explore multimodal approaches to address this gap. One promising direction is the integration of models like Audio2Face, which can simulate a human-like avatar of the person on the other end by analyzing vocal cues. This avatar would mirror user expressions, generate empathetic responses, and offer more emotionally aware interactions. Such a system could serve not only as a conversational partner but also as a source of emotional support, especially in applications like mental health, virtual companionship, or user engagement.

By combining audio, visual, and textual modalities, we hope to push the boundaries of emotionally intelligent AI systems and create more nuanced, responsive, and supportive human-AI interactions.

CONCLUSION

SoulSync demonstrates how modern AI techniques can bring accessible mental health support to users. By combining multimodal sensing with a retrieval-augmented language model, the system empowers users with early, judgment-free assistance. It is not meant to replace therapists but to act as a first line of support that encourages self-care and destigmatized help-seeking (aligning with our goals of reducing provider burden and improving emotional well-being). Our evaluation showed that even with an open-source backbone model, SoulSync produces coherent, relevant responses while offering rapid feedback. In the future, we will refine the emotion detection models, expand language support, and conduct broader user studies. We believe tools like SoulSync can make a tangible difference in people's lives by making emotional support more private, personal, and immediate.

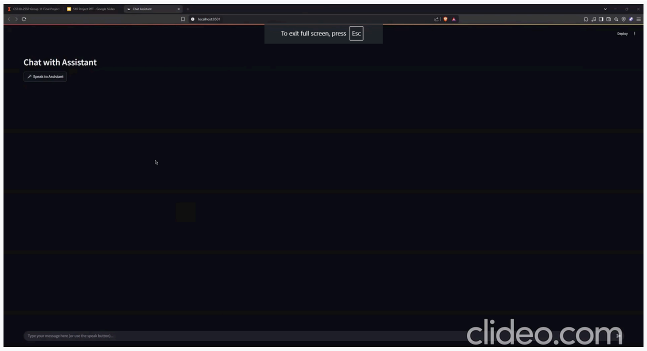
REFERENCES

1. https://www.cdc.gov/tobacco/campaign/tips/diseases/depression-anxiety.html#print
2. https://huggingface.co/OpenGVLab/InternVL2_5-1B
3. Chen, Z., Wang, W., Tian, H., Ye, S., Gao, Z., Cui, E., Wang, W. (2024). How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *Science China Information Sciences*, *67*(12), 220101.
4. Here's your citation in ACM conference format:
5. Gemini Team, Rohan Anil, Sebastian Borgeaud, and Jean. 2025. *Gemini: A Family of Highly Capable Multimodal Models.* In Proceedings of the ACM Conference on Computational Linguistics. arXiv:2312.11805.
6. https://elevenlabs.io/docs/capabilities/speech-to-text
7. https://www.pinecone.io/
8. https://build.nvidia.com/nvidia/audio2face-3d

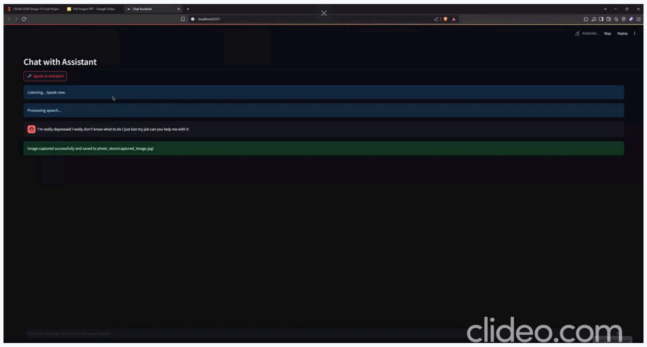
APPENDIX

Details about how to use SoulSync with a use case:

1. When the tool is running, users will see the chat option immediately in front of their screen. They can chat with the tool by typing their query (text) or capturing a live audio-video clip by clicking on 'Speak to Assistant' button.


CS510-25SP Group 11 Final Project
510 Project PPT - Google Slides
Chat Assistant
localhost:8501
To exit full screen, press Esc
Deploy
Chat with Assistant
Speak to Assistant
Type your message here (or use the speak button)...
clideo.com

2. When the user speaks with SoulSync, the tool will interact with the user in real time. It will do real time audio processing and capture video feed as well that will be used for facial expression detection. They can see all messages on the screen itself on the tool.

3. 

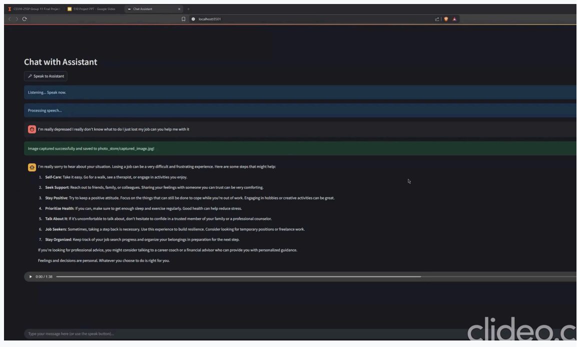
Listening... Speak now.

Processing speech...

I'm really depressed I really don't know what to do I just lost my job can you help me with it

Image captured successfully and saved to photo_store/captured_image.jpg!

4. Once the multimodal tool processes its inputs, it will give feedback on how to cope up with the user's issue in real time. You can also click the play button to listen to the generated response.


CS510-25SP Group 11 Final Project
510 Project PPT - Google Slides
Chat Assistant
localhost:8501
Speak to Assistant
0:00 / 1:38
Type your message here (or use the speak button)...
clideo.c

Chat with Assistant

Listening... Speak now.

Processing speech...

I'm really depressed I really don't know what to do I just lost my job can you help me with it

Image captured successfully and saved to photo_store/captured_image.jpg!

I'm really sorry to hear about your situation. Losing a job can be a very difficult and frustrating experience. Here are some steps that might help:

1. **Self-Care:** Take it easy. Go for a walk, see a therapist, or engage in activities you enjoy.
2. **Seek Support:** Reach out to friends, family, or colleagues. Sharing your feelings with someone you can trust can be very comforting.
3. **Stay Positive:** Try to keep a positive attitude. Focus on the things that can still be done to cope while you're out of work. Engaging in hobbies or creative activities can be great.
4. **Prioritize Health:** If you can, make sure to get enough sleep and exercise regularly. Good health can help reduce stress.
5. **Talk About It:** If it's uncomfortable to talk about, don't hesitate to confide in a trusted member of your family or a professional counselor.
6. **Job Seekers:** Sometimes, taking a step back is necessary. Use this experience to build resilience. Consider looking for temporary positions or freelance work.
7. **Stay Organized:** Keep track of your job search progress and organize your belongings in preparation for the next step.

If you're looking for professional advice, you might consider talking to a career coach or a financial advisor who can provide you with personalized guidance.

Feelings and decisions are personal. Whatever you choose to do is right for you.