

Multimodal retrieval-augmented generation with sentiment-aware responses for depression

Aluri Suguna Ratnapriya, A. Mary Sowjanya

Department of Computer Science and Systems Engineering, College of Engineering, Andhra University, Visakhapatnam, India

Article Info

Article history:

Received Jul 14, 2025

Revised Aug 4, 2026

Accepted Aug 12, 2026

Keywords:

Large language models

Mental health

Multimodal fusion

OpenAI

Prompt sentiment

Sentiment-aware responses

ABSTRACT

In view of the rapid growth in conversational interfaces, there is a demand for more adaptive retrieval systems. Traditional chatbots, built on pre-trained models, struggle to understand and respond to a wide range of real-world user queries. In view of this, a framework based on multimodal retrieval-augmented generation with sentiment-aware responses (MRAG-SAR) for depression that addresses the challenges and strengthens conversational interface performance has been developed. The proposed system can process unstructured data in multiple formats of inputs, such as audio, image, and text, using a multimodal late-fusion mechanism, making it suitable for managing diverse, domain-specific information. It also leverages powerful embedding models for capturing contextual similarity and scalable vector databases for efficient retrieval. The user queries are augmented to provide precise and nuanced context-related responses. The proposed framework retrieves content that captures the dimensions of the topic to enhance the relevance of the output responses. Introduction of sentiment-aware response generation allows the system to respond in a tone based on the emotional context of the user queries, which is essential in making use of personalized conversations. Relevant bidirectional encoder representations from transformers score (BERTScore) and other metrics highlight the framework's effectiveness in delivering high-quality responses.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Aluri Suguna Ratnapriya

Department of Computer Science and Systems Engineering, College of Engineering, Andhra University

Visakhapatnam, India

Email: sugunaratnapriya.rs@andhrauniversity.edu.in

1. INTRODUCTION

One of the most substantial concerns that has been affecting individuals of different age groups in modern society is mental health. A study of the increase in mental health disorders has indicated that mental well-being is related to productivity, social interactions, and overall life satisfaction. Mood disorders, anxiety disorders, and stress-induced disorders lead to a curtailed quality of life. Most people may not even know where to start or whom to discuss with when they experience mental health issues; addressing such mental health challenges has become critical. But the stigma associated with mental health makes individuals reluctant to reach out for help. In this context, artificial intelligence (AI) can play a transformative role in bridging the gap.

The latest technology involving AI-powered virtual assistants has shown great potential in delivering immediate support and appropriate care on demand. AI has aided valuable awareness in assessing mental health disorders [1]. Machine learning approaches enabled high classification accuracy, demonstrating the effectiveness in detecting emotional states from unstructured social media text; it reinforced the importance of integrating contextual and semantic understanding in the automated analysis of

mental health signals [2]. Machine learning algorithms like support vector machine, gradient boosting machine, random forest, naive Bayes, and k-nearest neighbors have their own advantages and limitations, and the algorithm suitability depends on factors like data size, dimensionality, and ethical considerations in clinical contexts [3].

Machine learning-based approaches help detect stress from text messages on social media and emphasize the critical role of feature engineering, data diversity, and algorithmic suitability depending on input modality, besides identifying key challenges related to data privacy, generalizability, and the need for standardized evaluation protocols [4]. In an early multi-level classification framework for depression detection in the context of social media analytics, text-based features such as sentiment, frequency of negative words, and usage patterns have been found to be reliable indicators of mental health conditions [5]. Transformer-based chatbots helped achieve high accuracy in generating relevant responses corresponding to the query's emotional context [6].

Natural language processing advances in terms of large language models (LLMs), such as generative pre-trained transformer (GPT) and bidirectional encoder representations from transformers (BERT), trained on ample text data, allow them to comprehend human language. However, the LLMs are only general-purpose models trained on various news articles, websites, and books, and their applications in specialized domains such as mental health can be less than optimal. Companies build chatbots for organizational use and train these systems only with organizational data, such as customer service scripts, product-related frequently asked questions (FAQs), and other domain-specific issues. Chatbots provide answers and solutions based on a predefined knowledge base. Their scope is often limited to the company's business operations, and addresses general customer service queries that lack the depth and nuance required for addressing sensitive topics like mental health. Retrieval-augmented generation (RAG) based approaches have enhanced the accuracy of the responses embodying knowledge from external sources. Multimodal RAG is useful where different input forms, like images, text, and videos, are used.

This study, therefore, proposes an innovative approach using multimodal retrieval-augmented generation with sentiment-aware responses (MRAG-SAR) to bridge this gap. This hybrid model combines the power of information retrieval with generative language models to provide personalized, sentiment-aware responses. MRAG-SAR integrates image, audio, and textual modalities, which inherently qualify it as a multimodal AI system. Unlike standard RAG, MRAG-SAR prioritizes emotion as a primary signal to infer the user's state (e.g., sadness, anxiety, and crisis risk) and uses it to adapt the response tone (calm, validating, and hopeful). The user query is used for similarity-based retrieval, while its emotional tone is assessed separately to guide the final response tone and safety. Sentiment does not control document retrieval or ranking. This yields responses that feel right and stay safe, especially when the user's stability is uneven. This intertwines the affective computing goal of creating empathetic interactions between humans and machines. MRAG-SAR's sentiment-aware response generation is aligned with the broader objectives of intelligent healthcare systems, specializing in mental and behavioral well-being.

2. RELATED WORKS

Kamoji *et al.* [7] reviewed the potential of using both language models and LLMs to anticipate mental healthcare. Singh *et al.* [8] proposed a LangChain framework to build MindGuide, a chatbot that mimics a mental health therapist. LLM-based approaches help identify disparities and errors in the generated responses [9]. LLMs represent transformative advancements in AI, with broad applications across industries. Different LLMs have paved the way into real-world applications [10]. A GPT-based method for depression detection has been reported to identify depressive linguistic patterns and contextual cues from patient speech, which showcases the effectiveness of advanced LLMs in early depression detection and emphasizes the need for responsible and clinically aware AI integration [11]. LLMs like GPT-4 have shown their ability to provide consistent and accurate results [12]. Transformer-based models can analyze therapeutic narratives to identify emotional states, coping mechanisms, and underlying mental health conditions [13]. Though LLMs ace the generation of responses, they struggle with hallucinations and biases [14]. Unlike LLMs, RAG hallucinates less and generates more factual and diverse responses [15]. Gao *et al.* [16] presented a RAG-based framework for answering science exam questions that integrates a retrieval component that searches domain-specific knowledge bases and a generation module that constructs coherent and accurate answers from retrieved passages.

Ascorbe *et al.* [17] developed a tool that addresses critical global health challenges by integrating AI-driven retrieval and generation. Sree *et al.* [18] used RAG and LLM to integrate various data formats into a healthcare chatbot named MEDGPT. Multimodal RAG comprehends different data formats and generates content [19]. RAG methods integrate multimodal information, such as images, code snippets, tables, graphs, and audio, to enhance generative model performance [20]. Cederlund *et al.* [21] harnessed the power of LLM

with RAG to reduce technician effort and to improve support workflows in automating digital tasks. Bag *et al.* [22] highlighted a document image retrieval technique using spatial and semantic analysis to align images with their text. Ma *et al.* [23] proposed a new framework called the empathetic response generation framework using reinforcement learning (EmpRL) that uses reinforcement learning to generate empathetic dialogue responses that better match the empathy level of human target responses, addressing the limitation of prior methods that rely solely on maximum likelihood training. Afzal *et al.* [24] reviewed datasets, emotion models, and key applications across modalities and discussed current challenges and outlined future research directions for advancing affective computing systems.

Radanliev [25] presented a comprehensive AI ethical framework that embedded core principles of transparency, fairness, and privacy to address bias and accountability. Ali *et al.* [26] presented a comprehensive and critical survey of explainable artificial intelligence (XAI), systematically reviewing explanation methods across model-agnostic, model-specific, local, and global approaches. They analyzed how XAI contributes to trustworthy AI by linking explainability with fairness, accountability, robustness, and human understanding. Mohammadi *et al.* [27] conducted a review on the use of AI in multimodal learning analytics, analyzing how data from multiple modalities are combined to model learning processes. Bevilacqua *et al.* [28] conducted a review of human-centered artificial intelligence (HCAI) to prioritize human needs, capabilities, and trust. They systematically searched major databases such as PubMed, ScienceDirect, IEEE Xplore, and analyzed how HCAI approaches are defined and applied.

Existing mental health chatbots lack a multimodal understanding and emotional alignment, which limits their effectiveness in sensitive contexts. Further, although many studies have focused on multimodal and context-aware RAG, to the best of our knowledge, no one has focused on sentiment-aware RAG. Such an approach uses multimodal inputs and multimodal fusion with emotional sensitivity for mental health support.

3. METHOD

The methodology of MRAG-SAR for depression is carefully crafted to understand and address questions about depression and mental health with both precision and empathy. A schematic representation of the MRAG-SAR is shown in Figure 1. The MRAG-SAR framework integrates multimodal retrieval, late fusion, and sentiment-aware response generation into a single processing pipeline.

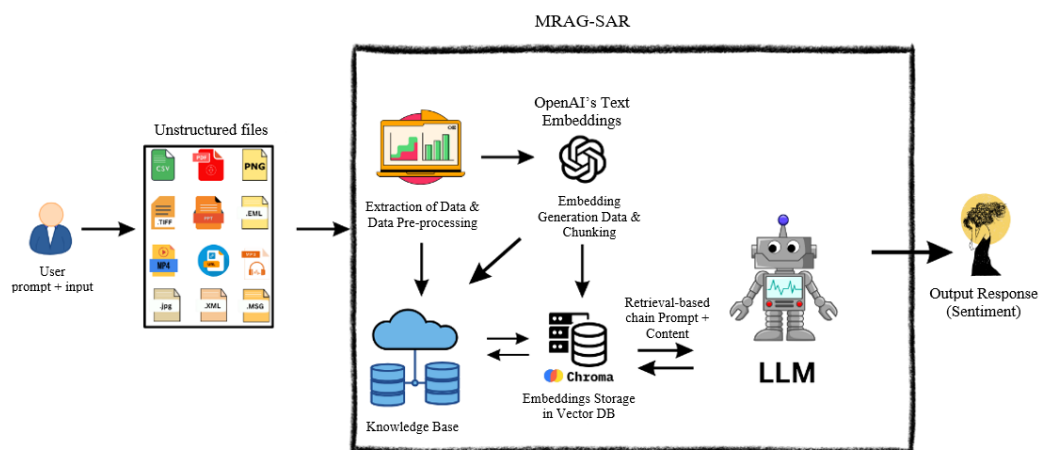


Figure 1. Schematic representation of the model developed in the present study

In multimodal fusion, text and image vectors exist in slightly different spaces, so that scores are not comparable. Simple concatenation or naive ranking can bury the most helpful evidence. To overcome these challenges, OpenAI embeddings are used for text, and bootstrapping language image pre-training (BLIP) representations are used for images, with the retrieved results integrated through late fusion. The system maintains separate text and image indexes, with audio transcripts included in the text index; the retrieved results were merged, tagged by modality, deduplicated by source, and reranked by similarity score. The query is represented appropriately for retrieval across modalities.

In emotional sensitivity, binary sentiment misses nuance (e.g., sadness, anxiety, and anger). Tone must adapt in real time without sounding scripted. To overcome these challenges, a sentiment-adaptive layer that consists of three steps was added: detection, assessment, and adaptation. Crisis routing uses

urgency thresholds and supportive escalation messages. In factual grounding, hallucinations may occur when the context is thin or conflicting, while mixed-quality sources may reduce accuracy and cause facts to become outdated as health guidance changes. To overcome these challenges, MRAG-SAR uses a curated domain-specific knowledge base and similarity-based retrieval to provide relevant evidence. Prompt-level guardrails are applied to keep the generated responses grounded in the retrieved context, safe, and contextually appropriate. Essential features of the methodology are outlined as follows.

3.1. Input

3.1.1. Text

As shown in Algorithm 1, MRAG-SAR can ingest and process various formats of mental health-related content, including PDFs and CSVs, by harnessing unstructured data loaders from a LangChain-supported library named unstructured. This flexible ingestion process enables heterogeneous depression-related resources to be incorporated into the knowledge base. It allows an assorted and extensive knowledge base that can be updated adaptively and customized to comply with the requirements of people seeking mental health support, capturing multiple perspectives and experiences that span academic research, clinical insights, and personal narratives.

Algorithm 1. MRAG-SAR for depression

```
MRAG-SAR (Q, D)
1: Input:   Q: User Query (e.g., "What are the signs of depression?")
           D: Set of unstructured documents D = {d1, d2, ..., dn} (e.g., documents: audio,
           text, image)
2: Output: A: Final sentiment-aware response to the query.
3: Procedure:
    3.1 Preprocessing and embedding:
        Preprocess: For each document d ∈ D, perform necessary preprocessing to handle
                     multimodal content.
        Embeddings: Compute the embedding for each document
                     V(d) = Embed(d)
                     Where Embed(*) is an embedding function.
    3.2 Vector store construction:
        Store Documents: Insert document d and its embedding V(d) into a vector database
                        VDB (e.g., ChromaDB, FAISS)
        Similarity Search Support: The vector database VDB enables similarity-based
                                   retrieval using metrics such as cosine similarity:

$$\text{Cosine}(V(Q), V(d)) = \frac{V(Q) \cdot V(d)}{\|V(Q)\| \times \|V(d)\|}$$

    3.3 Query and retrieval:
        Embed Query: Compute the embedding of the user query Q:
                     V(Q) = Embed(Q)
        Retrieve Relevant Documents: Retrieve the top K documents most similar to Q from
                                   VDB.

$$\Delta = \text{RetrieveTopK}(V(Q), \{V(d) \mid d \in D\})$$

        Where Δ is a subset of D, and it denotes the set of retrieved documents.
        Late Fusion: Merge and rerank combined multimodal results into a single context for
                     the LLM
        Augment Query: Construct a prompt P by combining Q with the content from the
                     retrieved set Δ:

$$P = \text{Prompt}(Q, \Delta)$$

    3.4 Model's response generation:
        Generate Answer: Pass the prompt P to the language model (LLM) to generate the final
        answer:

$$A = \text{LLM}(P)$$

        Return Final Output: Return A as the final sentiment-aware response to the query Q.
```

3.1.2. Audio

OpenAI's whisper model is used for audio files, which can transcribe voice recordings into text. Distress analysis interview corpus-wizard of oz (DAIC-WOZ) is a multimodal dataset for psychological distress analysis. The dataset comprises audio recordings, text transcripts, and metadata from semi-structured clinical interviews in which participants interact with a virtual agent. The dataset contains annotations of depression severity based on the patient health questionnaire (PHQ-8), enabling binary and severity classification tasks on this dataset. The speech audio is provided in WAV format [29], [30].

3.1.3. Image

The BLIP algorithm handles images (JPEG and TIFF), while optical character recognition (OCR) is used for text-rich images. Sometimes, individuals with depression may find it easier to express their feelings by speaking rather than writing. BLIP generates captions for images through a multi-stage process that

combines vision-language encoding and language generation, and also generates image captions by integrating a vision encoder and a language decoder within a unified architecture [31].

3.2. Vectorization

In MRAG-SAR, OpenAI embeddings are used for text, and BLIP embeddings for images. Embedding models convert textual data into dense vector representations, which aid in semantic search tasks. The use of ChromaDB in MRAG-SAR enables efficient storage, retrieval, and management of embeddings.

3.3. Augmentation

RAG is an advanced AI technique that retrieves relevant information or documents from a knowledge base in response to a user's query, and then LLM, such as GPT-4o, generates responses based on both the user query and the retrieved information. A modular architecture is used to enable independent processing, integration, and scalability across all system components. Figure 2 depicts a multimodal data ingestion and embedding pipeline for multimodal RAG. The input layer accepts PDFs, images, audio, and document uploads. The file processor stores files and routes them based on type—audio, image, or document. Each file type enters its own extraction branch using tools such as whisper (audio), OCR (image), or unstructured (documents). Extracted text is chunked into smaller segments for embedding generation. Text embeddings are created using OpenAI embeddings; image embeddings via BLIP multimodal encoders. Extracted numeric facts (like PHQ-9 scores) are processed separately. ChromaDB stores text and image embeddings for retrieval; SQLite stores structured facts. This framework supports multimodal retrieval and semantic search for later use in RAG. Overall, the system provides a unified backend for cross-modal knowledge indexing and retrieval.

The primary distinction between the present approach and traditional organizational chatbots is in the specialization of the data. While general-purpose LLMs are trained on vast amounts of data, including organizational content, the proposed system builds a rich, domain-specific knowledge base tailored to mental health, especially depression. Retrieval of contextually relevant content from various sources on mental health disorders is done by the proposed system, ensuring that the user receives not only generic information but also tailored, evidence-based support, rather than relying on a static set of FAQs or predefined scripts. MRAG-SAR's multimodal aspect allows diverse data sources such as text files, images, and audio, enhancing the depth of understanding, as well as enriching the response generation process by producing responses that are grounded in factual data and making the interaction more supportive and compassionate. Integrating RAG-based models in mental health applications enhances the quality and relevance of responses, particularly for those who may not have immediate access to professional help.

In the proposed system, multimodal fusion takes three different kinds of data: text, images, and audio, and processes each in the best way for that modality, then combines the results at the end. Text from documents, OCR, and audio transcripts is chunked and embedded with an OpenAI embeddings model into a text index; images are embedded visually with BLIP into a separate image index, and any text inside images is also OCR-recognized and fed into the text index. At query processing time, the user's question is embedded once with the text model to search all text-based content, and once with the BLIP text encoder to search the image embeddings in the same semantic space. This gives two sets of hits: text hits and image hits. These results are merged and tagged by modality, deduplicated by source to prevent a single file from dominating, and reranked by similarity score to produce a unified ranked list of the most relevant content across all modalities. The final, fused set of results is passed to the LLM to generate an answer; though each modality is handled separately during indexing and retrieval, the fusion occurs at the decision layer through a late-fusion approach.

3.4. Sentiment-aware response generation and model adaptation

Building on the retrieved context and embeddings, the system leverages GPT-4o, configured with targeted prompt templates, to generate coherent, contextually grounded responses. GPT-4o synthesizes factual excerpts from the knowledge base into empathetic, clear, and user-focused answers by structuring prompts with explicit instructions for tone, style, and safety. These prompt templates guide the model to balance factual grounding, emotional sensitivity, and response safety.

The novel concept of prompt sentiment is also included. The emotional tone of the prompt sentiment, which is embedded in user input, works as a key factor in influencing the behavior of LLMs. Positive, neutral, or negative sentiment in prompts affects LLM outputs in terms of factual accuracy, coherence, bias, and length. Sentiment-aware prompt engineering [32] is supported to ensure fairness and factual integrity in sensitive applications such as policy analysis, mental health chatbots, and legal or financial documentation. To align the model's behavior with mental health best practices, the authors incorporated domain-specific knowledge from a curated corpus of clinical guidelines, peer-reviewed studies, and anonymized patient narratives.

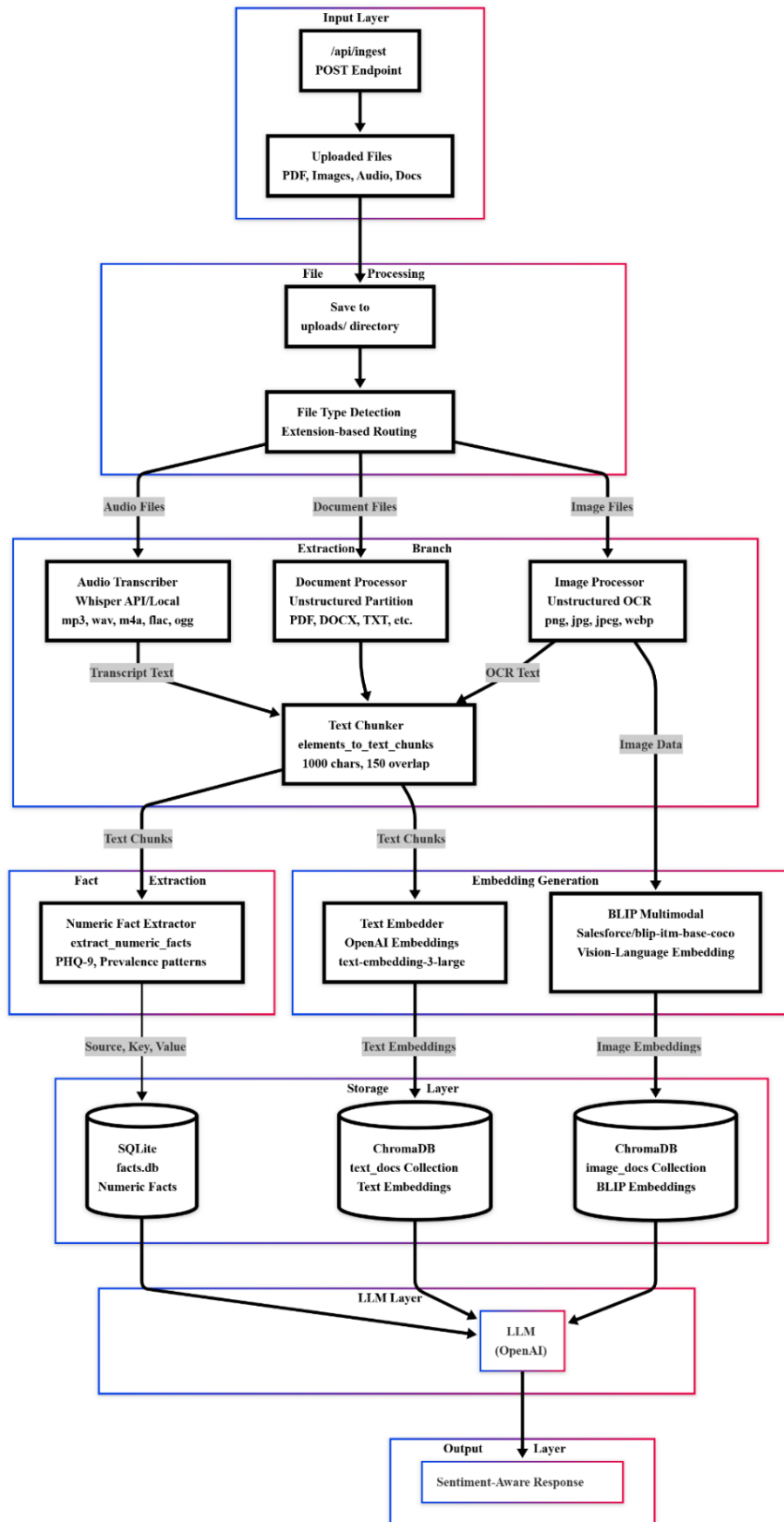


Figure 2. Modular architecture of the proposed MRAG-SAR

The domain-specific knowledge and prompt guidance improve the system's sensitivity to the nuances of depression-related language, helping it identify high-risk cues through prompt-level assessment and apply safety-oriented guardrails. Together, these measures ensure that every generated response is information-rich, accurate, supportive, respectful, and appropriate for individuals seeking mental health guidance. Many RAG systems lose emotional context during response generation. But, MRAG-SAR incorporates sentiment-aware response generation, ensuring the generation of emotionally aligned responses based on the retrieved evidence. To generate sentiment-aware responses, multiple guardrails have been used to ensure that responses generated by the MRAG-SAR remain safe, unbiased, and contextually appropriate.

4. RESULTS AND DISCUSSION

The retrieval mechanism, built on OpenAI embeddings, BLIP embeddings, and cosine similarity, enables the MRAG-SAR to identify highly relevant documents from the knowledge base for any query. The proposed MRAG-SAR not only detects sentiment but also dynamically adjusts tone and phrasing in mental-health contexts through the integration of a sentiment layer. With high compassion, accuracy, relevance, empathy, therapeutic appropriateness (CARET) scores, MRAG-SAR demonstrates promising performance in HCAI for emotionally delicate domains. It is deeply aligned with the broader goals of affective, explainable, and ethical AI, which are the three pillars shaping the next generation of emotionally intelligent systems. From an XAI standpoint, MRAG-SAR uses a modular architecture that supports component-level interpretability, with whisper for audio, BLIP for images, and GPT-4o for response generation. In terms of ethical AI, the system prioritizes user well-being and safety. MRAG-SAR ensures that personal data is stored minimally as required, as data privacy is crucial in healthcare. From an affective AI standpoint, MRAG-SAR demonstrates the ability to understand and respond to human emotions, supported by multimodal emotion recognition across text, audio, and images.

Figure 3 illustrates the image caption generation by BLIP when presented with an image of a person wearing a white shirt and smiling. Figure 4 illustrates the image caption generation by BLIP when presented with an image of a child crying. Figure 5 demonstrates the conversion of audio to text by OpenAI whisper. The audio file was taken from the DAIC-WOZ dataset of participant 301.

Quality of the generated responses was evaluated using BERTScore, a text generation metric that calculates the similarity between a generated text and a reference text. Instead of relying on exact token matches, BERTScore leverages contextual embeddings to capture the semantic similarity between the texts. Comparison of bidirectional encoder representations from transformers score (BERTScores) obtained with different inputs for LLM (GPT-4o) and MRAG-SAR indicated higher BERTScores for MRAG-SAR compared to LLM (GPT-4o).

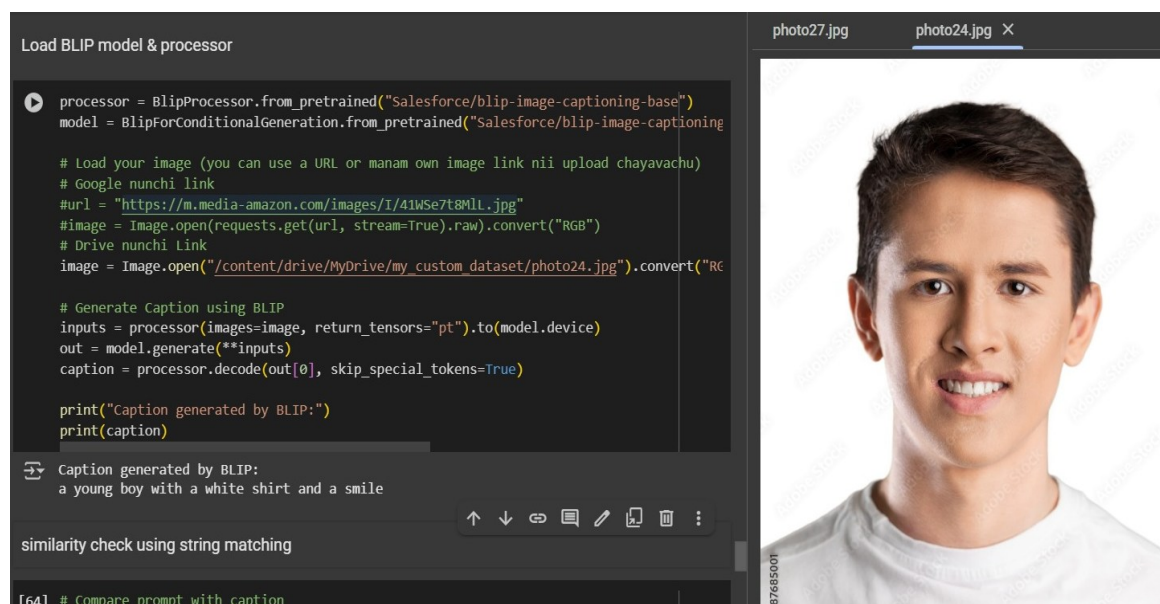
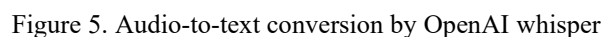
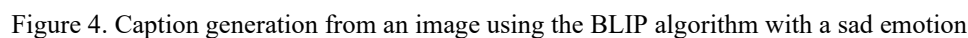


Figure 3. Caption generation from an image using the BLIP algorithm with a smile emotion



Int J Artif Intell, Vol. 15, No. 5, October 2026: 4907-4918

the baseline LLM. The user-study results further demonstrated high empathy and emotional alignment for MRAG-SAR. The BLIP multimodal encoder and whisper support modality-specific image and audio processing, respectively, within MRAG-SAR's modular response-generation pipeline. Sentiment prompting pays attention to sentiment dynamically, modifying prompts in real-time based on detected emotions, while achieving alignment accuracy and maintaining high empathy and compassion. Since rule-based systems work differently, as they are based on hard-coded rules, while the RAG system depends on LLM reasoning, we have not compared them.

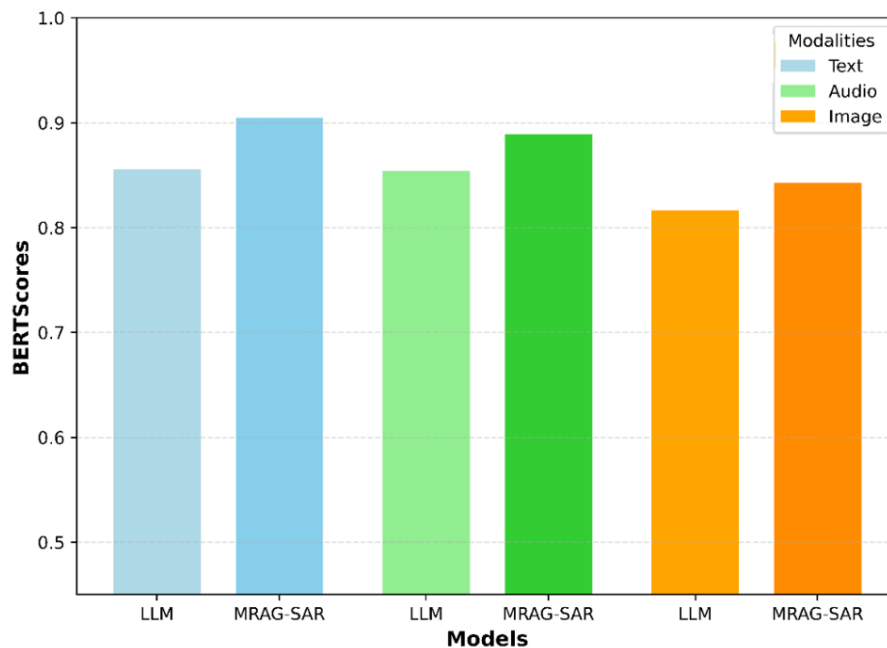


Figure 6. Comparison of BERTScores for modalities

A real-world user study was conducted involving 35 participants across diverse age groups from 16 to 62 years, comprising students, software professionals, and a senior citizen. The survey collected self-reported emotional states, stress triggers, and behavioral responses to evaluate how well the MRAG-SAR system resonates with authentic human sentiment profiles. The model's generated responses were then evaluated for empathy, emotional alignment, and supportive tone. Participants rated these responses on a CARET-based scale, while one domain expert later reviewed a sample of anonymized conversations for clinical sensitivity.

As depicted in Figure 7, key insights from the user study included an average CARET score of 89.7%, along with emotional alignment of 97.1%, response safety of 90.3%, crisis detection of 85.1%, sentiment alignment accuracy of 88.0%, and factual consistency scores of 87.4%, which are higher than the values obtained for GPT-4 and standard RAG, except for response safety, which was higher in GPT-4 than in MRAG-SAR. This finding is consistent with GPT-4's broader general-purpose safety alignment and policy safeguards, whereas MRAG-SAR is configured primarily for depression-specific retrieval and prompt guidance. MRAG-SAR's sentiment-aware response generation is aligned with the broader objectives of intelligent healthcare systems, specializing in mental and behavioral well-being. Younger users valued supportive tone and validation, while older participants emphasized clarity and reassurance. Career pressure, relationship issues, family expectations, financial concerns, and grief-related stress were some of the common stress factors identified in responses, which are aligned with current mental-health literature, indicating ecological validity. The emotional alignment achieved by MRAG-SAR is consistent with prior studies showing that generative chatbots and empathy-aligned response generation can enhance contextual and emotional support in mental-health conversations [6], [23].

A structured error analysis is conducted on MRAG-SAR to interpret emergent behaviors and identify improvement areas across its multimodal reasoning and sentiment-aware generation framework. MRAG-SAR exhibited strong domain-specific performance in mental health contexts. Because MRAG-SAR was optimized for emotional alignment, prioritizing emotional understanding, which led to better HCAI, it

yielded higher sentiment alignment accuracy of 88.0%. Similarly, the factual consistency obtained is also higher compared to that obtained from GPT-4 and standard RAG. Since the knowledge base is strictly confined to the healthcare domain, the system may not retrieve relevant documents for general queries. This may result in vague responses, and hallucinations are possible. MRAG-SAR, being a domain-specific RAG model, may exhibit overfitting due to behavioral bias resulting from its adaptation to depression-related contexts. MRAG-SAR was evaluated using URL-based knowledge retrieval from the World Health Organization's depression-related records. These records represent a diverse and internationally balanced group of countries and cultural contexts and are not confined to a single regional background.

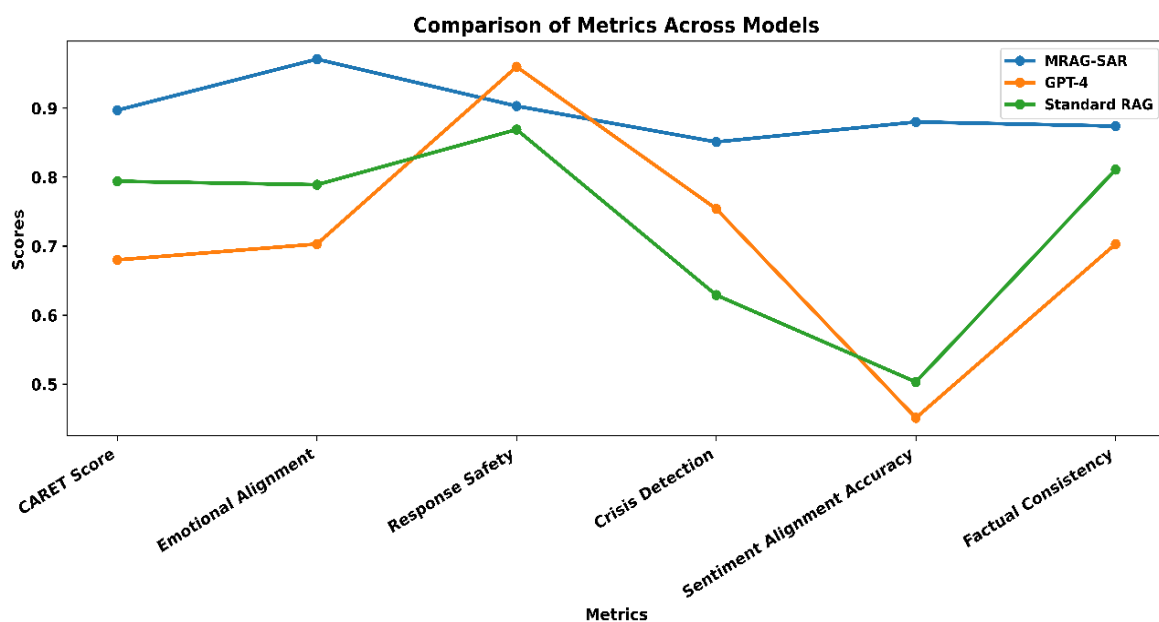


Figure 7. Comparative evaluation of MRAG-SAR with GPT-4 and standard RAG

5. CONCLUSION

The obtained results clearly demonstrated that the MRAG-SAR for depression effectively handled depression-related queries. Integrating RAG enriched with a sentiment layer enabled the system to produce sentiment-aware responses that are empathetic and emotionally aligned. In contrast to a purely task-oriented approach, MRAG-SAR demonstrates how generative systems can be laced with compassion to provide safe, supportive, accessible, and empathetic responses for underprivileged populations experiencing emotional distress. The system's multimodal design enhances accessibility for users with diverse communication preferences. The MRAG-SAR architecture is also adaptable to further multimodal extensions, such as videos, to provide enhanced support for mental health. MRAG-SAR can adopt federated fine-tuning for privacy-preserving mental health support, allowing local updates on user devices without transferring personal data. The core competency of MRAG-SAR is its modularity, which allows it to be adapted seamlessly across domains. The same sentiment modules can be extended to other mental health issues like anxiety and post-traumatic stress disorder (PTSD), to provide emotionally supportive interaction and also to support learners emotionally, adapting to multimodal communication. It can be fine-tuned on domain-specific data for emotionally intelligent AI across multiple sectors.

FUNDING INFORMATION

The authors state that no funding was received for this work.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Aluri Suguna Ratnapriya	✓	✓	✓	✓	✓	✓		✓	✓		✓		✓	
A. Mary Sowjanya		✓		✓		✓	✓	✓		✓		✓		✓

C : Conceptualization

M : Methodology

So : Software

Va : Validation

Fo : Formal analysis

I : Investigation

R : Resources

D : Data Curation

O : Writing - Original Draft

E : Writing - Review & Editing

Vi : Visualization

Su : Supervision

P : Project administration

Fu : Funding acquisition

CONFLICT OF INTEREST STATEMENT

The authors state no conflict of interest.

DATA AVAILABILITY

The data that support the findings of this study are available from <https://dcapswoz.ict.usc.edu/> and <https://www.gettyimages.in/>.

REFERENCES

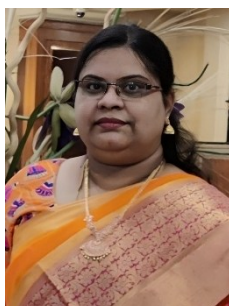
- [1] A. Mittal, L. Dumka, and L. Mohan, "A comprehensive review on the use of artificial intelligence in mental health care," in *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, Jul. 2023, pp. 1–5, doi: 10.1109/ICCCNT56998.2023.10308255.
- [2] T. Nijhawan, G. Attigeri, and T. Ananthakrishna, "Stress detection using natural language processing and machine learning over social interactions," *Journal of Big Data*, vol. 9, no. 1, Dec. 2022, doi: 10.1186/s40537-022-00575-6.
- [3] G. Cho, J. Yim, Y. Choi, J. Ko, and S.-H. Lee, "Review of machine learning algorithms for diagnosing mental illness," *Psychiatry Investigation*, vol. 16, no. 4, pp. 262–269, Apr. 2019, doi: 10.30773/pi.2018.12.21.2.
- [4] V. Arya and A. K. Mishra, "Machine learning approaches to mental stress detection: a review," *Annals of Optimization Theory and Practice*, vol. 4, no. 2, pp. 55–67, 2021, doi: 10.22121/aotp.2021.292083.1074.
- [5] M. M. Aldarwish and H. F. Ahmad, "Predicting depression levels using social media posts," in *2017 IEEE 13th International Symposium on Autonomous Decentralized System (ISADS)*, Mar. 2017, pp. 277–280, doi: 10.1109/ISADS.2017.41.
- [6] J. J. Bird and A. Lotfi, "Generative transformer chatbots for mental health support: a study on depression and anxiety," in *16th International Conference on Pervasive Technologies Related to Assistive Environments*, Jul. 2023, pp. 475–479, doi: 10.1145/3594806.3596520.
- [7] S. Kamaji, S. Rozario, S. Almeida, S. Patil, S. Patankar, and H. Pendhari, "Mental health prediction using machine learning models and large language model," in *2024 Second International Conference on Inventive Computing and Informatics (ICICI)*, Jun. 2024, pp. 185–190, doi: 10.1109/ICICI62254.2024.00040.
- [8] A. Singh, A. Ehtesham, S. Mahmud, and J.-H. Kim, "Revolutionizing mental health care through LangChain: a journey with a large language model," in *2024 IEEE 14th Annual Computing and Communication Workshop and Conference (CCWC)*, Jan. 2024, pp. 73–78, doi: 10.1109/CCWC60891.2024.10427865.
- [9] Z. Hammane, F.-E. B. -Bouazza, and A. Fennan, "SelfRewardRAG: enhancing medical reasoning with retrieval-augmented generation and self-evaluation in large language models," in *2024 International Conference on Intelligent Systems and Computer Vision (ISCV)*, May 2024, pp. 1–8, doi: 10.1109/ISCV60512.2024.10620139.
- [10] M. A. K. Raiaen *et al.*, "A review on large language models: architectures, applications, taxonomies, open issues and challenges," *IEEE Access*, vol. 12, pp. 26839–26874, 2024, doi: 10.1109/ACCESS.2024.3365742.
- [11] M. Danner *et al.*, "Advancing mental health diagnostics: GPT-based method for depression detection," in *2023 62nd Annual Conference of the Society of Instrument and Control Engineers (SICE)*, Sep. 2023, pp. 1290–1296, doi: 10.23919/SICE59929.2023.10354236.
- [12] G. Lorenzoni, P. E. Velmovitsky, P. Alencar, and D. Cowan, "GPT-4 on clinic depression assessment: an LLM-based pilot study," in *2024 IEEE International Conference on Big Data (BigData)*, Dec. 2024, pp. 5043–5049, doi: 10.1109/BigData62323.2024.10825184.
- [13] P. Bookanakere, S. Sharanya, S. Saniya, R. Arjun, and K. S. Srinivas, "Textual therapies: harnessing language models for mental health insights," in *2024 International Conference on Advances in Computing Research on Science Engineering and Technology (ACROSET)*, Sep. 2024, pp. 1–5, doi: 10.1109/ACROSET62108.2024.10743711.
- [14] Y. Chang *et al.*, "A survey on evaluation of large language models," *ACM Transactions on Intelligent Systems and Technology*, vol. 15, no. 3, pp. 1–45, Jun. 2024, doi: 10.1145/3641289.
- [15] P. Lewis *et al.*, "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *34th Conference on Neural Information Processing Systems (NeurIPS 2020)*, 2020, pp. 9459–9474.
- [16] M. Gao, J. Xiong, Y. Zong, and S. Li, "Science exam question answering based on retrieval-augmented generation," in *2024 3rd International Conference on Computer Technologies (ICCTech)*, Feb. 2024, pp. 55–58, doi: 10.1109/ICCTech61708.2024.00020.
- [17] P. Ascorbe, M. S. Campos, C. Domínguez, J. Heras, and A. R. T. -Reinares, "Towards a retrieval augmented generation system for information on suicide prevention," in *2023 IEEE EMBS Special Topic Conference on Data Science and Engineering in Healthcare, Medicine and Biology*, Dec. 2023, pp. 143–144, doi: 10.1109/IEEECONF58974.2023.10404508.
- [18] Y. B. Sree, A. Sathvik, D. S. H. Akshit, O. Kumar, and B. S. P. Rao, "Retrieval-augmented generation based large language model chatbot for improving diagnosis for physical and mental health," in *2024 6th International Conference on Electrical, Control and Instrumentation Engineering (ICECIE)*, Nov. 2024, pp. 1–8, doi: 10.1109/ICECIE63774.2024.10815693.

- [19] P. Joshi, A. Gupta, P. Kumar, and M. Sisodia, "Robust multi-model RAG pipeline for documents containing text, table & images," in *2024 3rd International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*, Jun. 2024, pp. 993–999, doi: 10.1109/ICAAIC60222.2024.10574972.
- [20] R. Zhao *et al.*, "Retrieving multimodal information for augmented generation: a survey," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023, pp. 4736–4756, doi: 10.18653/v1/2023.findings-emnlp.314.
- [21] O. Cederlund, S. Alawadi, and F. M. Awaysheh, "LLMRAG: an optimized digital support service using LLM and retrieval-augmented generation," in *2024 9th International Conference on Fog and Mobile Edge Computing (FMEC)*, Sep. 2024, pp. 54–62, doi: 10.1109/FMEC62297.2024.10710181.
- [22] S. Bag, A. Gupta, R. Kaushik, and C. Jain, "RAG beyond text: enhancing image retrieval in RAG systems," in *2024 International Conference on Electrical, Computer and Energy Technologies (ICECET)*, Jul. 2024, pp. 1–6, doi: 10.1109/ICECET61485.2024.10698598.
- [23] H. Ma, B. Zhang, B. Xu, J. Wang, H. Lin, and X. Sun, "Empathy level alignment via reinforcement learning for empathetic response generation," *IEEE Transactions on Affective Computing*, vol. 16, no. 3, pp. 1873–1884, Jul. 2025, doi: 10.1109/TAFFC.2025.3544594.
- [24] S. Afzal, H. A. Khan, M. J. Piran, and J. W. Lee, "A comprehensive survey on affective computing: challenges, trends, applications, and future directions," *IEEE Access*, vol. 12, pp. 96150–96168, 2024, doi: 10.1109/ACCESS.2024.3422480.
- [25] P. Radanliev, "AI ethics: integrating transparency, fairness, and privacy in AI development," *Applied Artificial Intelligence*, vol. 39, no. 1, Dec. 2025, doi: 10.1080/08839514.2025.2463722.
- [26] S. Ali *et al.*, "Explainable artificial intelligence (XAI): what we know and what is left to attain trustworthy artificial intelligence," *Information Fusion*, vol. 99, Nov. 2023, doi: 10.1016/j.inffus.2023.101805.
- [27] M. Mohammadi, E. Tajik, R. M. -Maldonado, S. Sadiq, W. Tomaszewski, and H. Khosravi, "Artificial intelligence in multimodal learning analytics: a systematic literature review," *Computers and Education: Artificial Intelligence*, vol. 8, Jun. 2025, doi: 10.1016/j.caeai.2025.100426.
- [28] R. Bevilacqua *et al.*, "Framing the human-centered artificial intelligence concepts and methods: scoping review," *JMIR Human Factors*, vol. 12, pp. e67350–e67350, May 2025, doi: 10.2196/67350.
- [29] J. Gratch *et al.*, "The distress analysis interview corpus of human and computer interviews," in *9th International Conference on Language Resources and Evaluation, LREC 2014*, May 2014, pp. 3123–3128, doi: 10.6331/73o7bccg9xequ.
- [30] D. DeVault *et al.*, "SimSensei kiosk: a virtual human interviewer for healthcare decision support," in *13th International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2014*, May 2014, pp. 1061–1068, doi: 10.65109/MXIV3169.
- [31] J. Li, D. Li, C. Xiong, and S. Hoi, "BLIP: bootstrapping language image pretraining for unified vision language understanding and generation," in *39th International Conference on Machine Learning*, 2022, pp. 12888–12900.
- [32] V. Gandhi and S. Gandhi, "Prompt sentiment: the catalyst for LLM change," in *2025 9th International Conference on Advances in Artificial Intelligence*, Nov. 2025, pp. 121–125, doi: 10.1145/3787279.3787299.

BIOGRAPHIES OF AUTHORS



Aluri Suguna Ratnapriya     is a research scholar in the College of Engineering, Andhra University, Visakhapatnam, India. She received the B.Tech. degree in Information Technology from Bapatla Engineering College and an M.Tech. degree in Computer Science and Engineering from the College of Engineering, Andhra University. Her areas of interest include artificial intelligence, deep learning, machine learning, and natural language processing. She can be contacted at email: priya3393@gmail.com.



A. Mary Sowjanya     is working as an associate professor in the Department of Computer Science and Systems Engineering, Andhra University. She did her B.Tech., M.Tech., and Ph.D. in Computer Science from Andhra University and has been a faculty member in the same department since 2006. She has four patents, more than 80 research publications to her credit, and received the "Young faculty research fellowship award" under the Visvesvaraya program by the Government of India. Her research interests include areas of data mining, machine learning, natural language processing, and data analytics. She can be contacted at email: dr.amsowjanya@andhrauniversity.edu.in.