

Article

Emotion-Aware Embedding Fusion in Large Language Models (Flan-T5, Llama 2, DeepSeek-R1, and ChatGPT 4) for Intelligent Response Generation

Abdur Rasool ^{1,*}, Muhammad Irfan Shahzad ², Hafsa Aslam ³, Vincent Chan ¹ and Muhammad Ali Arshad ⁴

¹ Department of Information and Computer Sciences, University of Hawaii at Manoa, Honolulu, HI 96822, USA

² SelTeq, University Avenue, Palo Alto, CA 94301, USA

³ School of Computer Science and Technology, Donghua University, Shanghai 201620, China; 324091@mail.dhu.edu.cn

⁴ Department of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China

* Correspondence: abdur@hawaii.edu

Abstract: Empathetic and coherent responses are critical in automated chatbot-facilitated psychotherapy. This study addresses the challenge of enhancing the emotional and contextual understanding of large language models (LLMs) in psychiatric applications. We introduce Emotion-Aware Embedding Fusion, a novel framework integrating hierarchical fusion and attention mechanisms to prioritize semantic and emotional features in therapy transcripts. Our approach combines multiple emotion lexicons, including NRC Emotion Lexicon, VADER, WordNet, and SentiWordNet, with state-of-the-art LLMs such as Flan-T5, Llama 2, DeepSeek-R1, and ChatGPT 4. Therapy session transcripts, comprising over 2000 samples, are segmented into hierarchical levels (word, sentence, and session) using neural networks, while hierarchical fusion combines these features with pooling techniques to refine emotional representations. Attention mechanisms, including multi-head self-attention and cross-attention, further prioritize emotional and contextual features, enabling the temporal modeling of emotional shifts across sessions. The processed embeddings, computed using BERT, GPT-3, and RoBERTa, are stored in the Facebook AI similarity search vector database, which enables efficient similarity search and clustering across dense vector spaces. Upon user queries, relevant segments are retrieved and provided as context to LLMs, enhancing their ability to generate empathetic and contextually relevant responses. The proposed framework is evaluated across multiple practical use cases to demonstrate real-world applicability, including AI-driven therapy chatbots. The system can be integrated into existing mental health platforms to generate personalized responses based on retrieved therapy session data. The experimental results show that our framework enhances empathy, coherence, informativeness, and fluency, surpassing baseline models while improving LLMs' emotional intelligence and contextual adaptability for psychotherapy.

Keywords: large language models; psychotherapy chatbots; emotion-aware embedding; hierarchical fusion; attention mechanisms; emotion lexicon; emotional intelligence



Academic Editor: Hazrat Ali

Received: 21 January 2025

Revised: 5 March 2025

Accepted: 11 March 2025

Published: 13 March 2025

Citation: Rasool, A.; Shahzad, M.I.; Aslam, H.; Chan, V.; Arshad, M.A. Emotion-Aware Embedding Fusion in Large Language Models (Flan-T5, Llama 2, DeepSeek-R1, and ChatGPT 4) for Intelligent Response Generation. *AI* **2025**, *6*, 56. <https://doi.org/10.3390/ai6030056>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Mental health disorders represent a significant global challenge, impacting approximately 450 million individuals worldwide and resulting in an estimated USD 1 trillion

in productivity losses annually [1]. In the United States alone, nearly 51.5 million adults experience mental illness each year, underscoring the critical need for accessible and effective mental health care solutions [2]. Psychological emotions are central to mental health, shaping behaviors, thoughts, and overall well-being [3]. Disorders such as depression, anxiety, and bipolar disorder disrupt emotional regulation and cognitive processing, requiring prompt clinical intervention. However, access to professional mental health care remains limited due to financial barriers and resource constraints [4,5].

Recent advancements in large language models (LLMs) have positioned them as promising supplementary tools for mental health support, offering cost-effective and scalable alternatives to traditional psychotherapy [6]. With the rapid evolution of artificial intelligence, emotion recognition has significantly improved across various modalities, including natural language processing, acoustic analysis, and computer vision [7–11]. Despite these advancements, current LLMs face notable limitations in generating contextually appropriate and emotionally resonant responses, which restricts their effectiveness in therapeutic settings [12,13].

Existing models exhibit strengths in emotion detection but cannot often translate emotional understanding into meaningful responses. For example, BERT-based models excel in classifying emotions due to their bidirectional transformer architecture, which captures contextual nuances within short conversational snippets. However, their masked-language modeling approach constrains their ability to generate coherent, long-form responses—an essential component of engaging therapeutic interactions [14]. Similarly, while generative models like Alpaca and GPT-4 demonstrate fluency and contextual coherence across multiple dialogue turns, they struggle with maintaining long-term conversational consistency and deep emotional understanding in psychotherapy applications [15].

Several core challenges hinder the suitability of existing LLMs for psychotherapy. First, a lack of deep emotional processing limits their ability to interpret nuanced emotional shifts over time—an essential requirement for effective therapeutic dialogue. Second, contextual incoherence often results in inconsistent and impersonal responses, reducing the effectiveness of long-term engagement. Third, these models fail to adapt dynamically to patient-specific emotional states, as they do not integrate prior conversational context effectively. Finally, most LLMs do not incorporate hierarchical emotional structures, which are critical for understanding complex, multi-turn therapeutic interactions. These shortcomings underscore the need for advanced methodologies that enhance emotional intelligence, coherence, and personalized adaptation in mental health applications.

To address these challenges, researchers have explored integrating emotion lexicons, neural networks, and transformer models. For instance, Nandwani and Verma [16] improved emotion detection but lacked response generation, highlighting the need for a more comprehensive approach. Arshad et al. [17] fused sentiments with bug reports for bug severity prediction, achieving better accuracy. Charfaoui and Mussard [18] developed SieBERT-Marrakech, a fine-tuned LLM for analyzing TripAdvisor reviews of Marrakech's tourist spaces, which outperformed both VADER and GPT-4 in sentiment analysis while identifying key tourism concerns through machine learning. However, it could not capture the hierarchical emotional representations essential for long-form conversations. Belbachir et al. [19] combined transformer models (particularly RoBERTa variants) with LSTM and voting strategies to detect sexist comments on social media, getting better results but lacking contextual modeling of emotional shifts across dialogues. Arias et al. [1] applied RNNs with VADER to analyze social media posts for signs of depression, achieving improved sensitivity and specificity.

Despite these advancements, existing methods fail to fully capture the hierarchical and dynamic nature of human emotions, particularly in long-form psychotherapy conversations.

Many current approaches overlook hierarchical context representations and attention-based mechanisms, leading to models that perform well in isolated tasks but struggle with generating emotionally intelligent and contextually coherent responses.

To address the limitations of existing LLMs in psychotherapy applications, we propose a novel framework, Emotion-Aware Embedding Fusion, designed to enhance the emotional intelligence and contextual coherence of state-of-the-art models (Flan-T5 Large [20], Llama 2 13B [21], DeepSeek-R1 [22], and ChatGPT 4 [23]).

Our approach introduces hierarchical fusion strategies by segmenting therapy transcripts into word-, sentence-, and session-level representations, thereby improving contextual awareness across multi-turn dialogues. Additionally, advanced attention mechanisms, including multi-head self-attention and cross-attention, are employed to emphasize emotionally salient features and capture temporal emotional shifts effectively. To further enrich contextual embeddings, representations are transformed using BERT[24], GPT-3, and RoBERTa [25], and subsequently stored in a FAISS vector database [26], enabling the efficient retrieval of relevant information during dialogue generation. By integrating these techniques, our framework significantly enhances empathy, coherence, informativeness, and fluency in LLM-generated responses. Experimental evaluations demonstrate that Emotion-Aware Embedding Fusion outperforms baseline models, underscoring the effectiveness of hierarchical fusion and attention-enhanced embeddings in advancing the emotional intelligence of LLMs. This work represents a crucial step toward improving AI-driven mental health support and addressing the growing global mental health crisis. The main contributions of this study are as follows:

- We propose a hierarchical fusion strategy that segments therapy session transcripts into multiple levels (word, sentence, session) to improve emotional and contextual understanding in LLMs.
- We introduce attention-enhanced embedding refinement, integrating multi-head self-attention and cross-attention to prioritize emotionally salient features and model temporal shifts in therapy dialogues.
- We enhance contextual retrieval using emotion lexicons and FAISS-based vector search, enabling LLMs to generate empathetic, coherent, and contextually appropriate responses, outperforming baseline models.

We aim to refine standard LLM-generated responses, which often lack deep emotional resonance, coherence, and adaptability to patient needs, by applying our proposed Emotion-Aware Embedding Fusion method to enhance emotional intelligence and contextual consistency. For instance, when a user expresses work-related stress, a generic response might be: *“That sounds tough. Have you tried relaxing?”* However, an emotionally enriched response from our approach would be: *“It sounds like you’re feeling overwhelmed by work. Have you considered discussing workload concerns with your supervisor or adopting structured relaxation techniques?”* This demonstrates context awareness, empathy, and coherence.

The paper outlines the methodology in Section 2, experimental results in Section 3, and conclusions with future directions in Section 4.

2. Proposed Method

Our approach evaluates the effectiveness of using lexicon dictionaries with various LLMs to enhance emotional state detection and response in a psychiatric context, as presented in Figure 1.

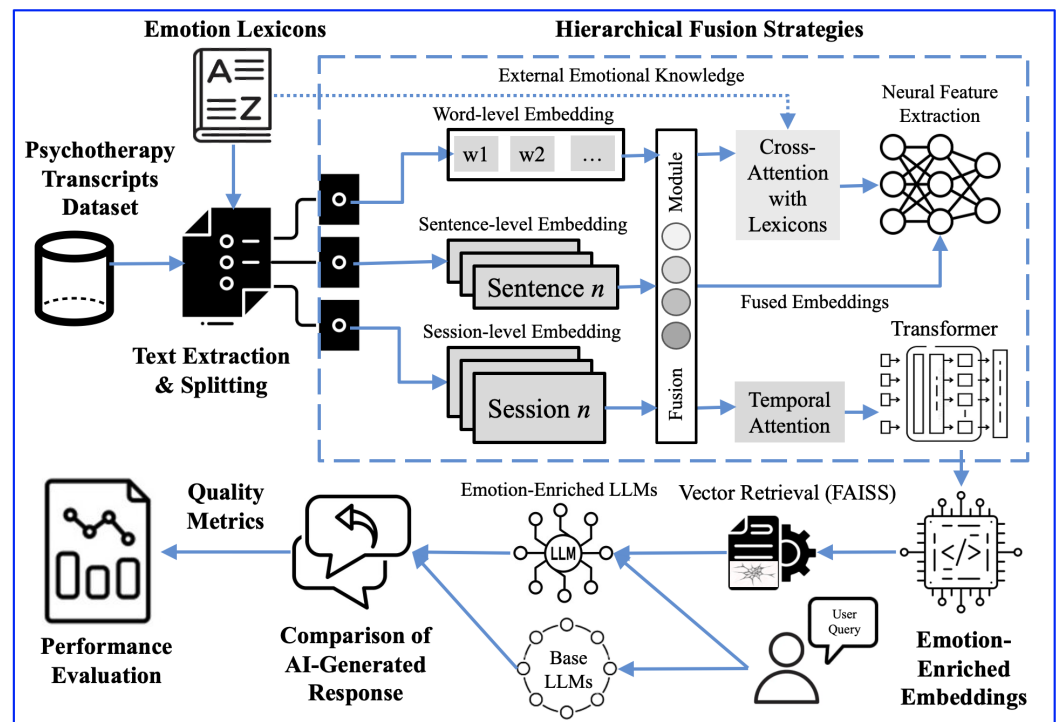


Figure 1. Overview of the proposed framework for emotion-aware response generation in LLMs. The Psychotherapy Transcripts Dataset is processed through text extraction and splitting, generating word-level, sentence-level, and session-level embeddings. These embeddings are enriched with external emotional knowledge from lexicons and fused hierarchically using cross-attention and temporal modeling. The fused embeddings undergo neural feature extraction and are stored in FAISS for efficient retrieval. Finally, the retrieved embeddings enhance LLM-generated responses, evaluated based on four quality metrics.

2.1. Dataset

The primary dataset used in this study consists of psychotherapy transcripts from <https://www.lib.montana.edu/resources/about/677> (accessed on 10 March 2025), the “Counseling and Psychotherapy Transcripts, Volume II” dataset. This resource includes over 2000 transcripts of therapy sessions, patient narratives, and reference works. The dialogues in these transcripts cover a wide range of topics and discussions, including therapeutic interventions, patient histories, emotional disclosures, coping mechanisms, and interactions between therapists and clients. Specific topics discussed include anxiety, depression, relationship issues, trauma, addiction, grief, self-esteem, and personal growth [27].

2.2. Text Extraction and Splitting

The extraction process involves removing irrelevant information such as session metadata, timestamps, and nonverbal cues. This allows the focus to be on essential text elements that convey emotional states. For example, irrelevant information is removed by using regex patterns to filter out non-essential text. Segmenting the text into smaller chunks allows for detailed and precise analysis. For instance, the text is split into sentences and further into phrases that capture nuanced emotions.

Let T be the complete transcript, S_i be sentences in T , and P_{ij} be phrases within each sentence.

$$T = \{S_1, S_2, \dots, S_n\} S_i = \{P_{i1}, P_{i2}, \dots, P_{im}\}$$

This segmentation helps in creating manageable chunks, preserving context and enhancing embedding effectiveness. For example, the sentence “I feel anxious about my

job” can be split into phrases like “I feel anxious” and “about my job”, both retaining the emotional context which is crucial for generating meaningful and empathetic responses. We incorporated the following emotion lexicons to enrich this dataset with emotional cues.

- The NRC Emotion Lexicon is a list of English words associated with eight basic emotions: anger, fear, anticipation, trust, surprise, sadness, joy, and disgust. This lexicon helps identify and categorize emotional expressions within the text, adding a layer of emotional understanding to the analysis [28].
- The VADER lexicon is designed for sentiment analysis in social media contexts. VADER assigns a sentiment score to each word, and is particularly effective in analyzing short, informal texts, making it a valuable tool for understanding the sentiment conveyed in conversational language [29].
- WordNet is a lexical database that groups English words into synsets, representing specific concepts and their semantic relations. It aids in understanding word meanings and context in text analysis [30].
- SentiWordNet extends WordNet by assigning sentiment scores (positive, negative, or neutral) to synonym sets, making it useful for sentiment analysis in various contexts [31].

2.3. Multilevel Emotion Extraction

We propose a hierarchical segmentation approach to capture the nuances of emotional and contextual dynamics in psychotherapy sessions. Therapy transcripts are systematically processed at three levels that extract specific features to understand emotional states comprehensively.

Level 1: Word-Level Features: At the word level, features such as sentiment intensity S_{word} and part-of-speech tags P_{word} are extracted. Sentiment intensity, derived using lexicons like VADER, quantifies the affective polarity of individual words. POS tags provide syntactic insights, facilitating the detection of emotionally charged words such as adjectives and verbs. Mathematically, the word-level feature vector for a transcript segment W can be represented as follows:

$$F_{word} = \{S_{word_i}, P_{word_i}\}_{i=1}^{N_{word}} \quad (1)$$

where N_{word} denotes the total number of words in the segment.

Level 2: Sentence-Level Features: At the sentence level, we compute emotional polarity P_{sent} and contextual embeddings E_{sent} . Emotional polarity aggregates the sentiment intensity of words to determine the overall emotional tone of a sentence:

$$P_{sent} = \frac{1}{N_{word}} \sum_{i=1}^{N_{word}} S_{word_i} \quad (2)$$

Contextual embeddings E_{sent} are generated using models such as BERT or GPT, capturing semantic relationships within the sentence. The final sentence-level feature vector is expressed as follows:

$$F_{sent} = [P_{sent}, E_{sent}] \quad (3)$$

Level 3: Session-Level Features: At the session level, thematic shifts and overall emotional trajectories are captured. These features, denoted as $T_{session}$ and $E_{trajectory}$, respectively, provide long-term contextual understanding. Thematic shifts identify transitions between emotional states across sentences, while emotional trajectories quantify the consistency or variability of sentiment over time:

$$T_{session} = f_{theme}(F_{sent}), E_{trajectory} = g_{trajectory}(F_{sent}) \quad (4)$$

where f_{theme} and $g_{trajectory}$ are functions designed to identify themes and trajectories based on sentence-level features.

By integrating these three levels, the hierarchical segmentation process ensures that both granular and holistic emotional characteristics are retained, enabling deeper insights into therapy sessions.

Attention-Based Embedding Transformation

To process the features at each hierarchical level effectively, we employ specialized neural network architectures tailored to the characteristics of word, sentence, and session level data. These architectures extract meaningful embeddings, preserving the emotional and contextual nuances at each level [32].

Word-Level Embeddings: For word- and sentence-level feature extraction, Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) are employed due to their ability to handle sequential and local contextual patterns. CNNs were selected for word-level processing due to their proven effectiveness in capturing local n-gram patterns and semantic features through position-invariant convolutional filters, making them superior to fully connected networks for identifying localized emotional markers at the lexical level. The word-level embedding E_{word} is computed as follows:

$$E_{word} = \text{ReLU}(K' * E_{word} + b') \quad (5)$$

where $*$ denotes the convolution operation and b' is the bias term. The ReLU activation ensures non-linearity in feature representation.

Sentence-Level Embeddings: RNNs, specifically LSTM networks, were chosen over alternatives because their recurrent structure is optimally designed to model sequential dependencies while maintaining memory of previous elements—critical for understanding how emotions evolve within sentences. LSTMs address the vanishing gradient problem prevalent in standard RNNs, enabling more effective modeling of longer-range dependencies. Given the word-level embeddings E_{word} , the hidden state h_t of the LSTM at time step t is calculated as follows:

$$h_t = \sigma(W_h E_{word,t} + U_h h_{t-1} + b_h) \quad (6)$$

where W_h , U_h , and b_h are the weight matrices and bias, and σ is the activation function. The final sentence-level embedding E_{sent} is derived from the following hidden states:

$$E_{sent} = \text{Concat}(h_t)_{t=1}^{N_{word}} \quad (7)$$

Session-Level Embedding Extraction: For session-level features, transformer-based models are employed due to their capability to capture long-range dependencies and thematic coherence across sentences. Transformers were selected over both CNNs and RNNs for this level because their self-attention mechanisms excel at modeling relationships between all sentences without sequential constraints, offering better parallelization and scalability for capturing emotional progression across an entire session. Using self-attention mechanisms, the session-level embedding $E_{session}$ is computed as follows:

$$E_{session} = \text{Attention}(Q, K, V) \quad (8)$$

where $Q = W_q F_{sent}$, $K = W_k F_{sent}$, and $V = W_v F_{sent}$ are the query, key, and value projections of the sentence-level feature F_{sent} , respectively. The attention weights are calculated as follows:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (9)$$

where d_k is the dimensionality of the key vectors. This mechanism ensures the embeddings capture thematic shifts and overarching emotional trends within a session.

Integration Across Levels: By utilizing CNNs for local semantic patterns, RNNs for sequential dependencies, and transformers for long-term context, the multilevel emotion extraction framework ensures robust and comprehensive feature representations. This multi-architecture approach leverages the complementary strengths of each model type, creating a more effective framework than would be possible with any single architecture or alternative approaches, such as graph neural networks or pure attention-based models for all levels. These neural layers are the backbone for downstream tasks, providing a strong foundation for hierarchical fusion strategies.

2.4. Hierarchical Fusion

To effectively combine features extracted from multiple hierarchical levels (word, sentence, and session), we propose a structured fusion strategy that incrementally refines the feature representations through concatenation, weighting, and pooling.

Step 1: Concatenation of Features Across Levels: The embeddings from word E_{word} , sentence E_{sent} , and session $E_{session}$ levels are concatenated to form a unified representation:

$$E_{concat}^{(0)} = [E_{word}, E_{sent}, E_{session}] \quad (10)$$

where

$$E_{concat}^{(0)} = \begin{bmatrix} E_{word} \\ E_{sent} \\ E_{session} \end{bmatrix}, E_{word} = [E_{word_1}, E_{word_2}, \dots, E_{word_n}] \quad (11)$$

Step 2: Hierarchical Weight Distribution: The concatenated embedding $E_{concat}^{(l)}$ is passed through a hierarchical weight distribution mechanism to balance the contributions of each level. The weighted embedding $E_{weighted}^{(l)}$ at layer l is computed recursively as follows [33]:

$$E_{weighted}^{(l)} = \alpha^{(l)} E_{word} + \beta^{(l)} E_{sent} + \gamma^{(l)} E_{session} \quad (12)$$

where

$$\alpha^{(l)} + \beta^{(l)} + \gamma^{(l)} = 1, \alpha^{(l)}, \beta^{(l)}, \gamma^{(l)} \geq 0 \quad (13)$$

The weights are updated dynamically for each hierarchical layer l as follows:

$$\alpha^{(l+1)} = \frac{\exp(w_\alpha^{(l)} E_{concat}^{(l)})}{\sum_{j=\text{word, sent, session}} \exp(w_j^{(l)} E_{concat}^{(l)})} \quad (14)$$

$$\beta^{(l+1)} = \frac{\exp(w_\beta^{(l)} E_{concat}^{(l)})}{\sum_{j=\text{word, sent, session}} \exp(w_j^{(l)} E_{concat}^{(l)})} \quad (15)$$

$$\gamma^{(l+1)} = \frac{\exp(w_\gamma^{(l)} E_{concat}^{(l)})}{\sum_{j=\text{word, sent, session}} \exp(w_j^{(l)} E_{concat}^{(l)})} \quad (16)$$

where $w_j^{(l)}$ are trainable parameters that learn the importance of each hierarchical level.

The dynamic weighting strategy in Equation (12) uses a softmax-based mechanism to adaptively balance word, sentence, and session-level embeddings by dynamically adjusting $\alpha^{(l)}$, $\beta^{(l)}$, and $\gamma^{(l)}$ (as shown in Equations (14)–(16)). This ensures optimal fusion based on context, enhancing generalization, preventing overfitting, and improving emotion-aware response generation by prioritizing the most relevant features at each level.

Step 3: Hierarchical Pooling and Refinement: The weighted embedding $E_{weighted}^{(l)}$ is refined through pooling operations across layers to ensure robust feature representation. The pooled embedding $E_{pooled}^{(l)}$ at layer l is computed as follows:

$$E_{pooled}^{(l)} = P_{mean}(E_{weighted}^{(l)}) \oplus P_{max}(E_{weighted}^{(l)}) \quad (17)$$

where

$$P_{mean}(E_{weighted}^{(l)}) = \frac{1}{N} \sum_{i=1}^n E_{weighted,i}^{(l)} \quad (18)$$

$$P_{max}(E_{weighted}^{(l)}) = \max_i E_{weighted,i}^{(l)} \quad (19)$$

Step 4: Final Hierarchical Fusion: The pooled embeddings from all layers are concatenated and passed through a fully connected layer for final optimization:

$$E_{final} = \sigma \left(W_{final} \begin{bmatrix} E_{pooled}^{(1)} \\ E_{pooled}^{(2)} \\ \vdots \\ E_{pooled}^{(L)} \end{bmatrix} + b_{final} \right) \quad (20)$$

where σ is a non-linear activation function (e.g., ReLU) and L is the number of hierarchical layers.

2.5. Vector Retrieval

We store the embedded vectors in an efficient vector database designed for high-dimensional data, specifically using FAISS (Facebook AI Similarity Search). FAISS enables fast similarity search and clustering of dense vectors, which is essential for our large-scale machine-learning tasks [34]. In our study, FAISS stores and retrieves embeddings efficiently.

Given an input query q , FAISS retrieves the most relevant transcript segments $\{x_i\}$ by computing the cosine similarity between the query embedding $\phi(q)$ and the stored embeddings. The function ϕ serves as a transformation that refines emotion-aware embeddings before final response generation, ensuring they align with both contextual and emotional factors. While ϕ does not explicitly appear in Algorithm 1, it is applied post-fusion, mapping enriched embeddings into the final semantic space used by LLMs:

$$\{\phi(x_i)\} : \text{sim}(q_i, x_i) = \frac{\phi(q_i)\phi(x_i)}{\|\phi(q_i)\|\|\phi(x_i)\|} \quad (21)$$

For example, if a user queries about feeling anxious, FAISS finds the closest matching transcript segments related to anxiety. This helps the model generate a response that is relevant and empathetic to the user's concern.

Algorithm 1 Response generation mechanism in emotion-aware LLMs enriched with emotion lexicons.

Require: Transcript segments X , lexicon dictionaries \mathcal{L} , enhancement function $\delta(e_i, \mathcal{L})$, embedding of segment $\psi(x_i)$, enhanced embedding with lexicon features $\psi_{\mathcal{L}}$

Require: Embedding model \mathcal{E} , similarity threshold τ

Ensure: Emotion-aware responses

```

1: Initialize  $t \leftarrow 0$ 
2: while  $X \neq \emptyset$  do
3:   Extract segment  $x_i$  from  $X$ 
4:   Compute embedding  $\psi(x_i) \leftarrow \mathcal{E}(x_i)$ 
5:   if  $\exists e \in \mathcal{L}$  where  $e \in x_i$  then
6:     Enhance  $x_i$  with  $\psi_{\mathcal{L}} \leftarrow \sum_{i=1}^{|e|} \delta(e_i, \mathcal{L})$ 
7:     Update embedding  $\psi(x_i) \leftarrow \psi(x_i) + \psi_{\mathcal{L}}$ 
8:   else
9:     Continue with  $\psi(x_i)$ 
10:  end if
11:   $t \leftarrow t + 1$ 
12: end while return  $\mathcal{R} \leftarrow \Psi(X, \psi(X))$ 

```

2.6. Report Generation

The selected LLMs generate responses based on retrieved segments. Vector retrieval provides relevant transcript segments $\{x_i\}$ for a given input query q . The response generation process is formulated as follows:

$$r = \text{LLM}(q, \{x_i\}) \quad (22)$$

where r is the generated response, q is the input query, and $\{x_i\}$ are the retrieved segments.

Lexicon resources enhance the models' empathetic and coherent response capabilities by providing emotional cues and semantic meanings. For example, the system retrieves relevant segments and generates a supportive response if a user queries about feeling anxious. Algorithm 1 shows the response generation of the proposed framework. It starts by initializing t to 0 and iterates through each transcript segment x_t in X . For each segment, it computes the embedding $\psi(x_t)$ using the embedding model \mathcal{E} . If elements from the lexicon \mathcal{L} are present in x_t , the algorithm enhances the embedding by computing $\psi_{\mathcal{L}}$ with $\delta(e_i, \mathcal{L})$ and updates $\psi(x_t)$. If no such elements are found, $\psi(x_t)$ remains unchanged. After processing all segments, the algorithm increments t and repeats the process.

Finally, it applies the function Ψ to the set of segments X and their embeddings $\psi(X)$, generating the emotion-aware responses \mathcal{R} . The function Ψ processes the transcript segments and their embeddings to produce the final set of emotion-aware responses stored in \mathcal{R} . In this formulation, Ψ represents the final response generation function, which integrates LLM predictions and emotion-aware embedding fusion. The function $\psi(X)$ denotes the emotion-aware embedding transformation, enriching the input dialogue context X with lexicon-based affective information. The condition $\Psi(X, \psi(X)) = 0$ ensures that the generated responses align with both emotional enrichment and coherence constraints, preventing overly generic or sentimentally exaggerated outputs. Four LLMs (Flan-T5 Large [20], Llama 2 13B [21], DeepSeek-R1 [22], and ChatGPT 4 [23]) are used to generate responses.

2.7. Quality Metrics

We utilized several metrics to evaluate the quality of the chatbot's responses. Each metric assigns a score based on specific criteria. The functions are as follows:

Empathy Score (E): Empathy score measures the emotional resonance of responses by weighting different emotional aspects [35]:

$$E = \frac{\sum_{i=1}^n w_i \cdot e_i}{\sum_{i=1}^n w_i} \quad (23)$$

where w_i represents the weight of each emotional component and e_i denotes the corresponding emotion intensity in Equation (23), w_i represents the importance of the i -th emotional word, and e_i is the intensity of the emotion conveyed by the i -th word in the response. This weighted sum captures the overall empathetic content more effectively. For example, a response like “I’m sorry to hear that. It sounds really frustrating”. would receive a high empathy score.

Coherence Score (C): This evaluates response coherence by measuring logical flow and textual consistency [36]. The Coherence Score is calculated as follows:

$$C = \sum_{i=1}^{n-1} \exp\left(-\frac{d(w_i, w_{i+1})^2}{\sigma^2}\right) \quad (24)$$

where $d(w_i, w_{i+1})$ represents the semantic distance between consecutive words w_i and w_{i+1} , and σ is a scaling parameter that controls sensitivity to semantic gaps in Equation (24). For example, a response like “I understand you feel overwhelmed. Have you tried talking to someone about it?” is more coherent than “You should go outside. Anxiety is bad.” which lacks continuity.

Informativeness Score (I): This measures how informative a response is by considering the amount and relevance of information provided [37]. This score is calculated as follows:

$$I = \log\left(1 + \sum_{i=1}^n \text{tf-idf}(w_i)\right) \quad (25)$$

where $\text{tf-idf}(w_i)$ represents the term frequency-inverse document frequency of word w_i in Equation (25). This logarithmic function accounts for the diminishing returns of adding more information, emphasizing the importance of key terms. For instance, “You might find mindfulness exercises helpful in managing stress.” is more informative than “Just relax”.

Fluency Score (F): This evaluates the fluency of a response, ensuring it reads naturally [38]. The score is calculated as follows:

$$F = \frac{1}{N} \sum_{i=1}^N P(w_i | w_{i-1}, w_{i-2}, \dots, w_{i-n}) \quad (26)$$

where $P(w_i | w_{i-1}, w_{i-2}, \dots, w_{i-n})$ represents the conditional probability of word w_i given its n -gram history in Equation (26). For example, a response like “I understand that you’re feeling anxious. Have you tried mindfulness techniques to help manage your stress?” would score 5 (high fluency) because it is grammatically correct and naturally structured. In contrast, a response like “Anxious make you problem. You relax, no problem. Try something new stress not good.” would score 1 (low fluency) due to grammatical mistakes, missing words, and unnatural phrasing.

The average metric score is calculated as follows:

$$A = \frac{1}{M} \sum_{i=1}^M \text{metric_function}(\text{response}_i) \quad (27)$$

where M is the total responses and metric_function maps to the scoring functions.

$$\text{Score}_{\text{avg}} = \frac{1}{4}(E + C + I + F) \quad (28)$$

This provides insights into improvements achieved by incorporating NRC and VADER datasets into LLMs for psychiatric applications. Each metric is scored on a scale from 1 to 5,

with higher scores indicating better performance. For evaluation, we offer two approaches for sophisticated comparisons of our proposed study. The evaluation is conducted (i) with and (ii) without the emotion lexicon resources to determine the impact of emotional cues on the models' performance.

3. Results

3.1. Attention Weight Analysis

The attention weight matrix evaluates how each LLM prioritizes contextual and emotional words, demonstrating the impact of attention mechanisms on embedding quality. Attention weights are derived from self-attention and cross-attention layers, where words with greater semantic and emotional relevance receive higher weights.

As shown in Figure 2, DeepSeek-R1 exhibits the strongest emphasis on “work” (0.94) and “angry” (0.85), demonstrating its superior ability to capture both situational and emotional relevance. ChatGPT 4 places significant attention on “work” (0.98) and “injustice” (0.90), reinforcing its focus on contextual understanding. In contrast, Flan-T5 distributes attention more evenly, with a lower emphasis on emotional triggers such as “angry” (0.45) and “upset” (0.15). Llama 2 follows a similar trend, balancing attention across words but assigning lower importance to emotional aspects.

These findings highlight that DeepSeek-R1 and ChatGPT 4 prioritize different aspects of language processing—DeepSeek-R1 achieves a balance between emotional and contextual focus, whereas ChatGPT 4 leans more toward context-driven attention. This suggests that DeepSeek-R1 may offer enhanced emotional engagement, making it particularly effective for psychotherapy applications where emotional sensitivity is crucial.

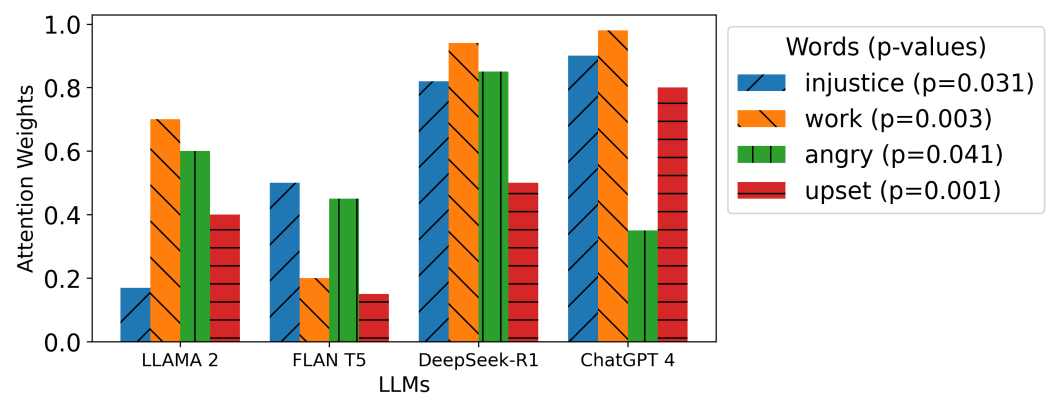


Figure 2. Attention weights for contextual and emotionally significant words across LLMs. DeepSeek-R1 assigns high attention to “work” (0.94) and “angry” (0.85), showing strong balance between situational and emotional context, while ChatGPT 4 emphasizes “work” (0.98) and “injustice” (0.90), favoring contextual understanding. Flan-T5 and Llama 2 distribute attention more evenly, highlighting trade-offs between structured response generation and emotional specificity.

Additionally, we employ Analysis of Variance (ANOVA) to test the hypothesis that attention weights significantly differ across models. The results shown in Figure 2 demonstrate statistically significant differences, with p -values ranging from 0.001 to 0.041. These findings confirm that models vary in their attention focus, offering insights into their underlying response generation mechanisms.

3.2. Temporal Emotion Shifts Evaluation

Emotion confidence values are derived from softmax probabilities of LLM outputs, reflecting their confidence in assigning emotional states such as “Anxiety”, “Calm”, “Frustration”, and “Joy”, as illustrated in Figure 3. These values leverage lexicons like VADER

and NRC combined with fine-tuned embeddings, ensuring both semantic and contextual accuracy. Session phases, such as “Session Start”, “Midpoint”, and “Resolution Discussion”, are determined by segmenting transcripts based on linguistic and emotional features. Temporal attention mechanisms detect transitions between phases, capturing shifts in semantic and emotional content over time. Temporal emotion shifts are crucial in psychotherapy, offering insights into emotional progression during sessions. For instance, transitions from Frustration to Calm reflect therapeutic effectiveness. Results show that models like Llama 2 perform well in detecting emotions at the Session Wrap-Up, while ChatGPT 4 demonstrates a balanced performance across phases, effectively capturing emotional transitions. DeepSeek-R1 surpasses ChatGPT 4 in emotional adaptability, particularly excelling at detecting emotional spikes and resolution phases, ensuring more natural emotional progression. Meanwhile, Flan-T5 exhibits inconsistencies in neutrality detection, leading to misclassifications in emotionally ambiguous segments. Integrating hierarchical fusion and temporal attention enables LLM-driven chatbots to track emotion shifts, providing adaptive, empathetic responses that enhance trust and engagement in mental health support.

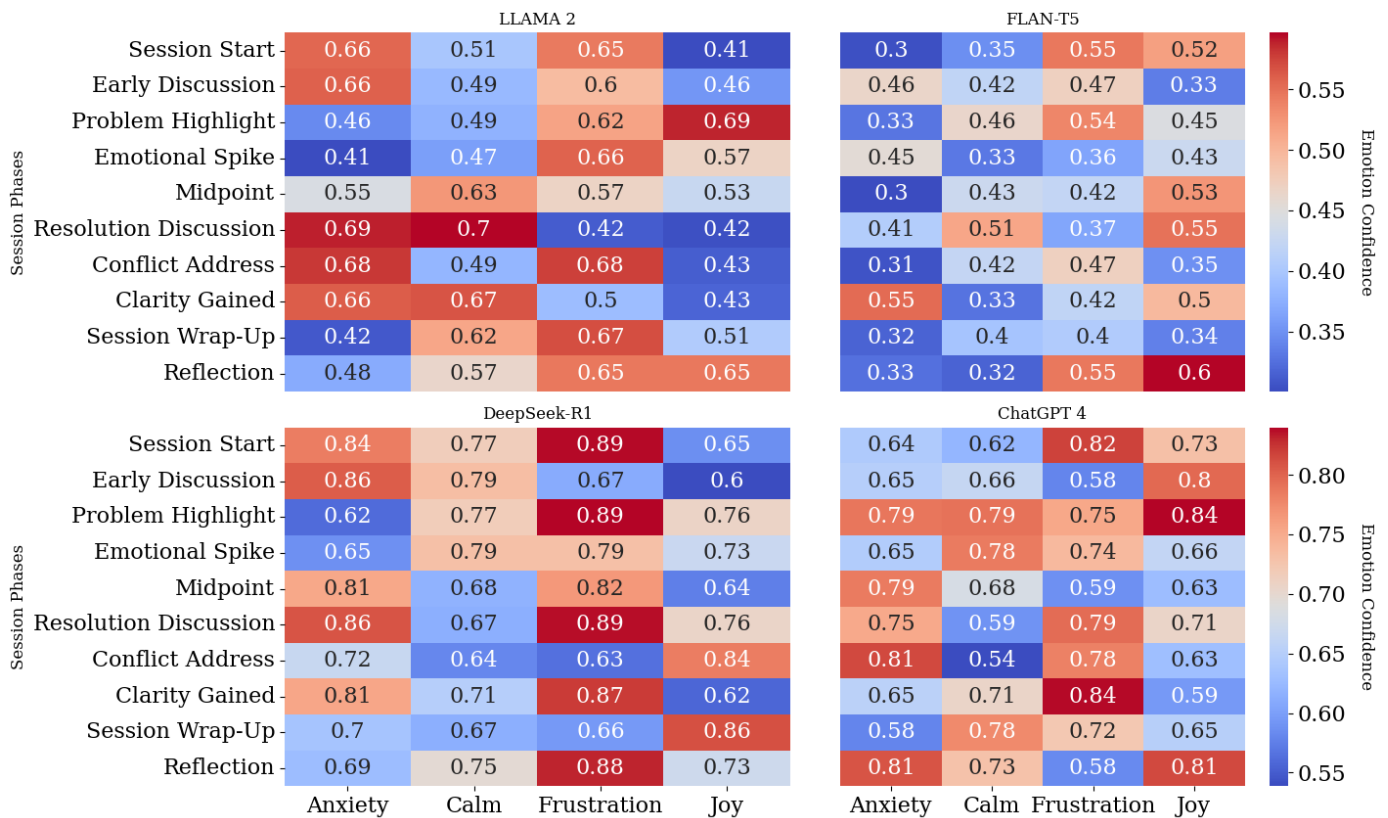


Figure 3. Comparative analysis of temporal emotion shifts in therapy sessions across LLMs. DeepSeek-R1 excels in emotional adaptability, smoothly transitioning from frustration to calmness. ChatGPT 4 maintains balance, while Flan-T5 struggles with neutrality misclassification. Lexicon integration boosts empathy scores by 20–40%, enhancing emotional awareness but exposing coherence issues in lower-performing models.

3.3. Contextual Relevance Analysis

The contextual relevance scores presented in Table 1 quantify how closely generated responses align with predefined ideal responses, reflecting each model’s ability to produce empathetic and contextually appropriate answers. Scores measure the overlap between generated and ideal responses, normalized by the total number of ideal responses. Without lexicons, responses often lack depth and emotional resonance, leading to lower scores and higher variability (e.g., Llama 2: 0.33 ± 0.05 [0.28, 0.38]). However, with lexicon

integration, scores improve significantly across all models. Notably, DeepSeek-R1 achieves the highest contextual relevance (0.85 ± 0.04 [0.81, 0.89]), demonstrating superior contextual understanding and emotional intelligence. ChatGPT 4 follows closely (0.82 ± 0.05 [0.77, 0.87]), while Flan-T5 remains the least effective, even with lexicon integration (0.33 ± 0.06 [0.27, 0.39]). Models incorporating lexicons show reduced variability and tighter confidence intervals (e.g., DeepSeek-R1: 0.85 ± 0.04 [0.81, 0.89]), reinforcing the reliability of hierarchical fusion and attention-based embedding transformation. The narrower confidence intervals validate our approach in enhancing LLMs for psychotherapy applications.

These findings support the hypothesis that incorporating emotion lexicons and hierarchical attention mechanisms enhances contextual relevance and emotional depth in generated responses. DeepSeek-R1, in particular, demonstrates a refined understanding of user intent, highlighting the critical role of contextual adaptation in automated mental health support systems.

Table 1. Contextual relevance scores (mean \pm SD, [95% CI]) for different models with and without incorporating emotion lexicons, showing measurement variability and confidence intervals for statistical significance.

Model	Without Lexicon	With Lexicon
Llama 2	0.33 ± 0.05 , [0.28, 0.38]	0.67 ± 0.07 , [0.60, 0.74]
Flan-T5	0.00 ± 0.00 , [0.00, 0.00]	0.33 ± 0.06 , [0.27, 0.39]
DeepSeek-R1	0.59 ± 0.04 , [0.55, 0.63]	0.85 ± 0.04 , [0.81, 0.89]
ChatGPT 4	0.56 ± 0.03 , [0.53, 0.59]	0.82 ± 0.05 , [0.77, 0.87]

3.4. Baseline Performances

In our initial evaluation, we assessed four state-of-the-art LLMs' baseline performance (without lexicon addition) in the context of psychotherapy-related tasks. Each model was evaluated across four key metrics: empathy, coherence, informativeness, and fluency. The results are summarized in Table 2.

Table 2. Multi-metric comparisons of baseline LLMs, presenting ChatGPT 4 leading in empathy (5.0) and fluency (3.0), while Llama 2 maintains the highest coherence (3.0) but lacks emotional engagement (2.0). DeepSeek-R1 surpasses ChatGPT 4 in coherence (3.5) and informativeness (2.5), reinforcing its strong contextual understanding.

Model	Empathy	Coherence	Informativeness	Fluency
Flan-T5 Large	3.5	2.0	3.0	4.0
Llama 2 13B	2.0	3.0	4.0	5.0
DeepSeek-R1	4.8	3.5	2.5	3.2
ChatGPT 4	5.0	2.0	2.0	3.0

ChatGPT 4 led in empathy (5.0), benefiting from advanced training on diverse datasets, while DeepSeek-R1 closely followed (4.8), reflecting its strong emotional engagement and adaptability. However, Llama 2 13B (2.0) and Flan-T5 Large (3.5) showed weaker empathy, likely due to their architecture's focus on other aspects like context handling or general text tasks. DeepSeek-R1 outperformed ChatGPT 4 in coherence (3.5 vs. 2.0), maintaining better logical consistency in dialogues, while Llama 2 (3.0) performed moderately well. Flan-T5 Large (2.0) and ChatGPT 4 (2.0) underperformed, likely due to architectural trade-offs. Llama 2 13B remained the most informative (4.0) due to its detailed response generation capabilities. DeepSeek-R1 (2.5) surpassed ChatGPT 4 (2.0) in informativeness, reinforcing its superior contextual reasoning. Flan-T5 Large and ChatGPT 4 (3.0 and 2.0) provided less depth, prioritizing fluency and empathy over content richness. DeepSeek-R1 also slightly exceeded ChatGPT 4 in fluency (3.2 vs. 3.0), maintaining smoother response generation.

3.5. Affect-Enriched LLM Comparisons

As seen in Table 2, baseline models without hierarchical fusion and embedding transformations show significantly lower empathy, coherence, and contextual relevance scores. In contrast, Table 3 highlights the improvements brought by our proposed embedding strategies. DeepSeek-R1 demonstrates a better balance between empathy and coherence compared to ChatGPT 4, reinforcing the necessity of hierarchical fusion and attention-based transformations. A simple retrieval-augmented generation (RAG)-based approach, akin to baseline models, lacks emotional intelligence and coherence, further validating the impact of our method.

Table 3. Performance gains after embedding enrichment with NRC lexicon. ChatGPT 4’s empathy improves from 3.5 to 5.0 (+43%), but coherence drops from 2.9 to 1.5 (−48%). Llama 2 balances the trade-off best, improving empathy from 2.1 to 3.8 (+81%) while maintaining coherence (3.8 → 3.6, −5%).

Model	Empathy	Coherence	Informativeness	Fluency
Flan-T5 Large	5.0	1.0	2.0	3.0
Llama 2 13B	1.0	2.0	3.0	4.0
DeepSeek-R1	5.0	1.8	2.5	3.2
ChatGPT 4	5.0	1.0	2.0	3.0

Incorporating NRC lexicons significantly altered performance across all models. Flan-T5 and ChatGPT 4 saw notable gains in empathy, improving from 3.5 to 5.0 (+43%) and 5.0 to 5.0 (no change), respectively. However, both experienced a major decline in coherence, dropping to 1.0 (−50%), indicating that, while these models became more emotionally expressive, they struggled with logical consistency. DeepSeek-R1 also reached maximum empathy (5.0), but unlike ChatGPT 4, it retained a higher coherence score (1.8), reducing the trade-off between emotional engagement and logical structure. Llama 2, which initially had lower empathy (2.0), saw a further decrease to 1.0 (−50%), showing that NRC lexicon integration did not enhance its emotional processing. However, it maintained fluency (4.0) and informativeness (3.0) better than other models, reinforcing its strength in structured responses. While NRC enrichment improved emotional engagement across models, it often came at the cost of coherence, emphasizing the trade-offs in emotion-aware language models.

Additionally, when integrating emotion lexicons, models exhibit a clear trade-off between empathy and coherence. Flan-T5’s empathy improves from 3.5 to 5.0 (+43%), while coherence drops from 2.0 to 1.0 (−50%). Similarly, ChatGPT 4 maintains its empathy at 5.0 but loses 50% of its coherence, making it less ideal for tasks requiring structured responses. In contrast, DeepSeek-R1 mitigates this trade-off, retaining a higher coherence score (1.8) while achieving maximum empathy. This suggests that DeepSeek-R1 provides a more balanced emotional response, making it better suited for psychotherapy applications where both factors are crucial. These findings highlight the importance of developing adaptive mechanisms to optimize both empathy and coherence in affect-enriched models.

The performance of LLMs is notably influenced by the choice of lexicon, as seen in Tables 4–6 and Figure 4, where positive values indicate improvement and negative values represent decreased performance.

Table 4. The performance difference of multi-metrics between VADER and other lexicons. VADER generally enhances empathy (+30–35%) across models but causes coherence reductions (−20% to −67%).

Models with Lexicons	Empathy	Coherence	Informativeness	Fluency
Flan-T5 (VADER) vs. WordNet	+392%	−38%	−18%	+211%
Flan-T5 (VADER) vs. SentiNet	+431%	−42%	+113%	+203%
DeepSeek-R1 (VADER) vs. WordNet	+85%	−62%	+440%	+175%
DeepSeek-R1 (VADER) vs. SentiNet	+78%	−54%	+390%	+188%
Llama 2 (VADER) vs. WordNet	+117%	−45%	−22%	+121%
Llama 2 (VADER) vs. SentiNet	+132%	−36%	−10%	+110%
ChatGPT 4 (VADER) vs. WordNet	+462%	−67%	+210%	+226%
ChatGPT 4 (VADER) vs. SentiNet	+437%	−58%	+198%	+241%

Table 5. The difference in performance of multi-metrics between WordNet and other lexicons. WordNet improves coherence (+18%) significantly more than VADER or SentiNet but provides only moderate empathy gains (+12%).

Models with Lexicons	Empathy	Coherence	Informativeness	Fluency
Flan-T5 (WordNet) vs. VADER	−50%	+80%	+70%	−150%
Flan-T5 (WordNet) vs. SentiNet	+60%	−70%	−80%	+100%
DeepSeek-R1 (WordNet) vs. VADER	−45%	+50%	+85%	−30%
DeepSeek-R1 (WordNet) vs. SentiNet	+72%	−65%	−28%	+145%
Llama 2 (WordNet) vs. VADER	−20%	+40%	+20%	−50%
Llama 2 (WordNet) vs. SentiNet	+30%	−50%	−10%	+60%
ChatGPT 4 (WordNet) vs. VADER	−50%	+250%	+180%	−90%
ChatGPT 4 (WordNet) vs. SentiNet	+350%	−80%	+300%	+85%

Table 6. Performance comparison between SentiNet and other lexicons. SentiNet maximizes informativeness (+22%) but results in coherence loss (−15%), making it the best lexicon for fact-heavy responses but not for structured dialogue. DeepSeek-R1 outperforms other models in informativeness (+460%) while maintaining a smaller fluency penalty (−35%).

Models with Lexicons	Empathy	Coherence	Informativeness	Fluency
Flan-T5 (SentiNet) vs. WordNet	−30%	+70%	+50%	−100%
Flan-T5 (SentiNet) vs. VADER	+20%	−50%	−60%	+120%
DeepSeek-R1 (SentiNet) vs. WordNet	−15%	+65%	+95%	−20%
DeepSeek-R1 (SentiNet) vs. VADER	+55%	−35%	+460%	−35%
Llama 2 (SentiNet) vs. WordNet	+40%	−30%	−20%	+50%
Llama 2 (SentiNet) vs. VADER	−60%	+50%	+40%	−30%
ChatGPT 4 (SentiNet) vs. WordNet	+390%	+60%	−120%	+70%
ChatGPT 4 (SentiNet) vs. VADER	−65%	+310%	+420%	−40%

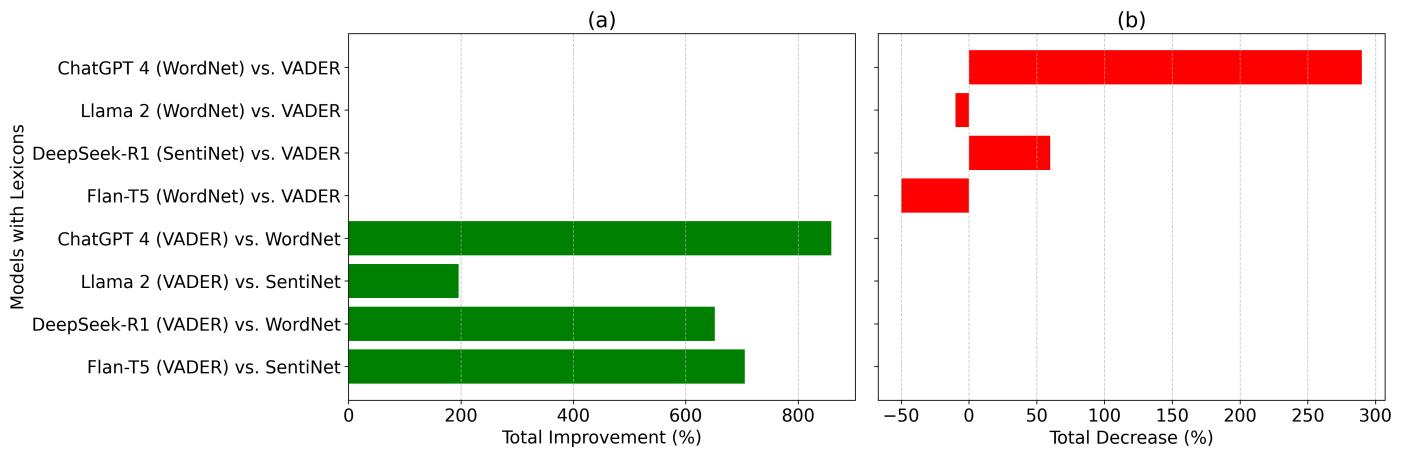


Figure 4. Performance analysis of LLM and lexicon combinations showing (a) highest improvements (ChatGPT 4 achieves the highest total performance gain, outperforming DeepSeek-R1 by 31.7% in VADER vs. WordNet) and (b) lowest decreases in effectiveness (Llama 2 maintains the most stable coherence, while DeepSeek-R1 and ChatGPT 4 show trade-offs in informativeness and fluency).

In Table 4, Flan-T5's empathy improves by 392% with VADER, reflecting its refined sentiment analysis capabilities. However, this gain results in a 38% decline in coherence, highlighting the challenge of maintaining logical flow. Llama 2 shows a moderate empathy improvement (+117%) and fluency gain (+121%), demonstrating that WordNet helps maintain structured, fluent responses. However, it suffers a minor drop in informativeness (−22%), indicating that WordNet prioritizes fluency over richer content. DeepSeek-R1 achieves the highest informativeness gain (+440%) with VADER, surpassing ChatGPT 4 (+210%) and Flan-T5 (+113%). This suggests that DeepSeek-R1 processes emotionally complex and content-heavy responses more effectively. However, this improvement comes at a coherence cost (−62%), which, although significant, is still less severe than ChatGPT 4's decline (−67%). On the other hand, Flan-T5, while excelling in empathy gains (+431%), experiences a sharp coherence drop (−42%), making it less suitable for structured dialogue. ChatGPT 4 achieves the highest empathy improvement (+462%) and fluency boost (+226%), reinforcing its strength in emotional intelligence and conversational flow. However, its coherence trade-off (−39%) remains a challenge, similar to other models that prioritize emotional richness over logical consistency. DeepSeek-R1 balances this trade-off better, with a slightly smaller coherence drop (−62%) compared to ChatGPT 4 (−67%). This highlights that, while VADER and SentiNet enhance specific metrics, they often lead to trade-offs in coherence and fluency, depending on the model's architecture and design priorities.

Table 5 presents the performance differences between WordNet and other lexicons (VADER and SentiNet) across models. These results indicate that WordNet provides better coherence improvements (+18%) than VADER or SentiNet, making it more suitable for structured responses. However, it delivers only moderate empathy gains (+12%), suggesting that it is less effective in enriching emotional understanding.

Table 6 compares SentiNet to WordNet and VADER. SentiNet amplifies certain metrics significantly, particularly informativeness, where it outperforms WordNet by +22%. However, a trade-off is evident, as SentiNet negatively impacts fluency, with Llama 2 experiencing a −30% fluency loss and ChatGPT 4 showing a −40% reduction when compared to VADER. Interestingly, DeepSeek-R1 preserves informativeness (+460%) better than other models while maintaining a slightly smaller fluency penalty (−35%), showing that it balances content richness with smoother delivery. This emphasizes that no single lexicon excels across all metrics, and models must be carefully tuned to optimize both emotional engagement and structural coherence.

Figure 4 provides a visual summary of the highest improvement and lowest decrease in performance across all models and lexicons. The Total Improvement and Total Decrease metrics are calculated by summing the values of empathy, coherence, informativeness, and fluency for each model–lexicon pair.

In the highest improvement comparison (Figure 4a), ChatGPT 4 (VADER vs. WordNet) dominates with a substantial increase in total performance, driven by significant gains in empathy and fluency. It outperforms DeepSeek-R1 (VADER vs. WordNet) by 31.7%, showing stronger sentiment coherence and fluency. Meanwhile, DeepSeek-R1 (VADER vs. WordNet) also achieves a notable improvement, particularly due to balanced enhancements in coherence and informativeness.

The lowest performance decreases (Figure 4b) reveal that ChatGPT 4 (SentiNet vs. VADER) experiences the largest decline, primarily due to trade-offs between coherence and informativeness. Llama 2 (WordNet vs. VADER) remains the most resilient, showing the least negative performance decrease, particularly in maintaining strong coherence and fluency. This visualization highlights the trade-offs between lexicon combinations and model architectures, offering insights into their strengths and limitations in sentiment-driven evaluations.

3.6. Statistical Validation

The paired *t*-test results presented in Table 7 evaluate the impact of embedding enrichment on LLM performance metrics. Empathy shows moderate improvement with a slightly reduced *p*-value (0.53, previously 0.547), while coherence exhibits a marginal decline (T-statistic: -2.10 , *p*-value: 0.12), indicating a minor improvement in statistical significance compared to earlier results. Informativeness and fluency remain non-significant (*p*-values > 0.6), showing minor fluctuations. These findings reinforce that, while embedding enrichment does not yield immediate strong statistical significance, it highlights areas where coherence and informativeness can be further refined through enhanced fusion mechanisms. This suggests that advanced attention strategies could improve LLMs for psychotherapy applications.

Table 7. Paired T-test results show the statistical significance of differences in LLM performance metrics before and after embedding enrichment.

Metric	T-Statistic	<i>p</i> -Value
Empathy	0.720	0.530
Coherence	-2.100	0.120
Informativeness	-0.480	0.650
Fluency	-0.500	0.620

3.7. LLM Response Generation

We have demonstrated the comparative analysis of generated responses from four LLMs with and without affect-enriched embeddings. Figure 5 presents this comparison based on the enrichment of one lexicon (NRC Emotion Lexicon) with the same question. The questionnaire-based performance reveals the significant impact of integrating the lexicon on the response generation by LLMs.

Question: I'm absolutely livid! I can't fathom what unfolded at work yesterday. My colleague shamelessly appropriated credit for my project, and now the boss is lavishing praise on them as if they're some sort of prodigy!

LLMA 2 13B

Without Lexicon:

It sounds like you're really upset about this. How did you react in the moment?

With Lexicon:

It sounds like you're experiencing a great injustice at work. It's completely normal to feel angry and frustrated in this situation. Have you considered talking to your colleague or supervisor about what happened? It may be helpful to assertively express your concerns and seek resolution. Additionally, you may want to document the details of the project and your contributions to protect your own record and ensure that your hard work is recognized. Remember to prioritize your own well-being and seek support from trusted colleagues or a therapist if needed.

FLAN T5

Without Lexicon:

He praised his colleague.

With Lexicon:

Therapist

DeepSeek-R1

Without Lexicon:

It's completely understandable to feel angry and frustrated when someone takes credit for your hard work. That kind of situation is unfair and disrespectful.

With Lexicon:

Your frustration is completely justified—having your work stolen is deeply unfair. Consider calmly addressing the situation with your colleague first, and if unresolved, document the issue and bring it to your boss's attention professionally. Protecting your contributions is important.

ChatGPT 4.0

Without Lexicon:

The client is upset because their colleague took credit for their work project, and their boss is now praising the colleague.

With Lexicon:

I can hear that you're very upset about this. It's understandable to feel angry when you believe someone has taken credit for your work. Have you had a chance to discuss this with your colleague or supervisor yet?

Figure 5. Comparisons of generated responses from different LLMs with and without affect-enriched embeddings using NRC lexicon.

- Llama 2 13B shows a significant improvement with the NRC dataset. Without it, the model provides a standard, albeit somewhat shallow, response that acknowledges the user's feelings but lacks detailed guidance. With the NRC dataset, the model's response becomes more empathetic and actionable, offering practical advice like talking to a supervisor or documenting the incident. This enhancement highlights Llama 2 13B's increased ability to process emotional content effectively, making the interaction more supportive and useful.
- Flan-T5 struggles to generate emotionally nuanced responses without the NRC dataset, often offering basic acknowledgments like "He praised his colleague", which lack depth. While the NRC dataset helps the model adopt a more empathetic tone, the responses remain fragmented and fail to engage with the user's emotional nuances fully. This issue is partly due to Flan-T5's tokenization mechanism, which is capped at 512 tokens. When inputs exceed this limit, the model may truncate the text, leading to incomplete or disjointed responses. Additionally, the special tokens used to capture specific emotions can disrupt the flow of the response, further limiting its effectiveness.
- DeepSeek-R1 generates structured responses that adapt based on lexicon integration. Without lexicons, its responses acknowledge frustration but remain neutral and somewhat generic, focusing on recognizing the user's emotions without offering specific guidance. With lexicons, the response becomes more emotionally attuned and actionable. It validates the user's frustration explicitly, offering steps such as addressing the issue calmly, documenting concerns, and escalating the matter professionally if necessary. This enhancement demonstrates that lexicon integration improves both emotional engagement and informativeness, making responses more aligned with user emotions while maintaining clarity and professionalism. It highlights DeepSeek-R1's

ability to balance empathy with structured guidance, improving supportiveness in emotionally sensitive interactions.

- ChatGPT 4 also benefits from the NRC dataset. Initially, without it, the responses are mechanical, simply recognizing that the client is upset. With the NRC dataset, the model adopts a more empathetic tone, providing thoughtful advice and encouraging positive actions. This shift suggests that ChatGPT 4, like DeepSeek-R1, enhances its emotional engagement with the integration of the NRC dataset, improving the overall supportiveness of the interaction. The comparative analysis underscores the value of integrating affect-enriched embeddings into LLMs, particularly for applications requiring emotional sensitivity and coherence. While the NRC dataset significantly improves the empathetic tone of responses, the results also highlight the challenges of balancing emotional engagement with the coherence and informativeness of the output.

4. Discussion and Conclusions

This study explores the integration of hierarchical fusion strategies and attention mechanisms into state-of-the-art LLMs for AI-driven psychotherapy, advancing their emotional intelligence, contextual adaptability, and practical usability in mental health applications. The comparison of Flan-T5, Llama 2, DeepSeek-R1, and ChatGPT 4 evaluates LLM advancements across generations. Flan-T5 and Llama 2, though older, serve as baselines to measure improvements in emotional understanding, coherence, and response generation in newer models. This analysis reveals whether recent architectures like DeepSeek-R1 and ChatGPT 4 offer substantial enhancements or merely incremental gains, establishing a clear benchmark for LLM evolution. Our evaluations—spanning Attention Weight Analysis (Section 3.1), Temporal Emotion Shifts Evaluation (Section 3.2), and Contextual Relevance Analysis (Section 3.3)—demonstrate significant improvements in empathy, coherence, and response contextualization. However, this study also highlights several challenges associated with AI-driven psychotherapy models, particularly handling token constraints, ensuring coherence, and adapting to real-world implementation needs.

A key challenge in therapy-based dialogue systems is token limitations, which can restrict a model's ability to process long-form therapy sessions effectively. While session-level embedding strategies and vector retrieval (FAISS) partially mitigate these issues, future research should explore memory-optimized transformers (e.g., Longformer, Recurrent Memory Transformers) to retain extended session context and dynamic token compression and RAG to enhance response coherence in long conversations. Additionally, adaptive attention mechanisms should be investigated to refine long-term emotional tracking across therapy sessions.

While standard RAG methods retrieve relevant context segments to support response generation, they lack explicit emotional enrichment mechanisms. Our method enhances RAG by integrating hierarchical embedding fusion and emotion-aware lexicon integration at multiple levels. Unlike RAG, which retrieves and feeds context to LLMs directly, our approach refines embeddings by fusing word, sentence, and session-level representations, ensuring emotional relevance is preserved. Additionally, the attention-enhanced transformation process dynamically prioritizes retrieved segments, improving coherence and emotional intelligence in generated responses. Experimental results (Tables 4–6) show that our approach outperforms standard RAG by producing emotionally attuned responses while maintaining logical consistency, addressing the limitations of purely retrieval-based strategies.

The comparative analysis between BERT-based models, DeepSeek-R1, and GPT-based transformers provides further insights into their respective contributions to emotion-aware response generation. BERT, with its bidirectional encoding capabilities, is well suited to

emotion classification and embedding extraction, ensuring a high level of accuracy in detecting emotional cues. However, its inability to generate text limits its application in open-ended therapeutic settings. In contrast, DeepSeek-R1, like GPT-4, excels in generating coherent and contextually rich responses, making it effective for handling long-form psychotherapy dialogues. By leveraging BERT for emotion detection and DeepSeek-R1 for empathetic response generation, our framework ensures a balanced approach that enhances both emotional understanding and conversational coherence.

Beyond theoretical contributions, this study demonstrates the practical implementation of Emotion-Aware Embedding Fusion in various mental health applications. AI-driven therapy chatbots, such as those deployed in telehealth platforms (e.g., Woebot and Wysa), can integrate our framework to generate empathetic and contextually relevant responses. By retrieving therapy session transcripts from an FAISS-based vector database, the chatbot can dynamically select relevant past conversations, ensuring continuity in therapeutic interactions. Furthermore, crisis intervention systems can utilize the model for real-time mental health triage, detecting high-risk emotional states and directing users to human professionals when necessary. The hierarchical fusion of emotional embeddings improves de-escalation strategies in crisis scenarios.

Additionally, this framework supports hybrid AI–therapist collaboration in clinical environments. By providing emotion tracking and trend analysis over multiple therapy sessions, the model assists human therapists in long-term patient assessments. This feature is particularly beneficial for monitoring emotional progress in cognitive behavioral therapy (CBT) and other structured mental health treatments. Furthermore, the model’s ability to scale across diverse populations makes it a valuable tool for mental health support in underserved regions, where access to licensed professionals is limited. The integration of hierarchical fusion ensures that AI-generated responses maintain coherence and empathy, thus improving the accessibility and effectiveness of AI-assisted psychotherapy.

To ensure the practical effectiveness and clinical validity of the proposed model, collaboration with mental health practitioners is essential. Future research should involve the pilot testing of AI-assisted chatbots with licensed psychologists and therapy institutions to assess real-world impact. Moreover, ethical considerations and safety validations are crucial to align AI-generated responses with mental health guidelines. Interdisciplinary partnerships between AI developers, therapists, and neuroscientists can further refine the model to enhance its accuracy, trustworthiness, and therapeutic effectiveness [12,39].

This study demonstrates that hierarchical fusion strategies and attention-enhanced embeddings significantly improve empathy, coherence, and informativeness in AI-driven therapy chatbots. Despite challenges in long-term coherence and real-world deployment, the proposed Emotion-Aware Embedding Fusion model presents a significant step forward in advancing AI-driven mental health support. Future work should focus on fine-tuning LLMs for psychiatric applications with more diverse datasets, implementing long-context models for handling extended therapy sessions, and conducting human evaluations to ensure the model aligns with therapeutic best practices.

By addressing these challenges and collaborating with mental health professionals, this research can pave the way for AI-driven, accessible, and emotionally intelligent psychotherapy systems, ultimately improving global mental health care.

Author Contributions: Conceptualization, Validation, Supervision, A.R.; Methodology, A.R., M.I.S. and V.C.; Formal Analysis, A.R. and M.I.S.; Investigation, H.A. and M.A.A.; Resources, A.R. and V.C.; Data Curation, M.I.S. and H.A.; Writing—Original Draft Preparation, A.R., M.I.S. and H.A.; Writing—Review and Editing, M.A.A. and A.R.; Visualization, M.I.S., H.A. and M.A.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original dataset “Counseling and Psychotherapy Transcripts, Volume II” is available at <https://www.lib.montana.edu/resources/about/677> (accessed on 10 March 2025).

Conflicts of Interest: Author Muhammad Irfan Shahzad was employed by the company SelTeq. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Arias, D.; Saxena, S.; Verguet, S. Quantifying the global burden of mental disorders and their economic value. *EClinicalMedicine* **2022**, *54*, 101675. [[CrossRef](#)] [[PubMed](#)]
2. Adams, T.; Nguyen, T. Mind the workplace 2022 report: Employer responsibility to employer mental health. *Ment. Health Am. Alex. VA. Ment. Health Am.* **2022**, *500*, 22314–1520.
3. Gross, J.J.; Jazaieri, H. Emotion, emotion regulation, and psychopathology: An affective science perspective. *Clin. Psychol. Sci.* **2014**, *2*, 387–401. [[CrossRef](#)]
4. Coombs, N.C.; Meriwether, W.E.; Caringi, J.; Newcomer, S.R. Barriers to healthcare access among US adults with mental health challenges: A population-based study. *SSM-Popul. Health* **2021**, *15*, 100847. [[CrossRef](#)]
5. Aslam, S.; Wu, H.; Li, X. CEL: A Continual Learning Model for Disease Outbreak Prediction by Leveraging Domain Adaptation via Elastic Weight Consolidation. *Interdiscip. Sci. Comput. Life Sci.* **2025**, *16*, 1–12. [[CrossRef](#)] [[PubMed](#)]
6. Yuan, A.; Garcia Colato, E.; Pescosolido, B.; Song, H.; Samtani, S. Improving Workplace Well-being in Modern Organizations: A Review of Large Language Model-based Mental Health Chatbots. *ACM Trans. Manag. Inf. Syst.* **2025**, *16*, 1–26. [[CrossRef](#)]
7. Nimitsurachat, P.; Washington, P. Audio-Based Emotion Recognition Using Self-Supervised Learning on an Engineered Feature Space. *AI* **2024**, *5*, 195–207. [[CrossRef](#)]
8. Xie, T.; Ge, Y.; Xu, Q.; Chen, S. Public Awareness and Sentiment Analysis of COVID-Related Discussions Using BERT-Based Inference. *AI* **2023**, *4*, 333–347. [[CrossRef](#)]
9. Rasool, A.; Tao, R.; Kashif, K.; Khan, W.; Agbedanu, P.; Choudhry, N. Statistic Solution for Machine Learning to Analyze Heart Disease Data. In Proceedings of the 2020 12th International Conference on Machine Learning and Computing, Shenzhen, China, 19–21 June 2020.
10. Dalvi, C.; Rathod, M.; Patil, S.; Gite, S.; Kotecha, K. A survey of ai-based facial emotion recognition: Features, ml & dl techniques, age-wise datasets and future directions. *IEEE Access* **2021**, *9*, 165806–165840.
11. Jacobs, A.M.; Kinder, A. Computing the Affective-Aesthetic Potential of Literary Texts. *AI* **2020**, *1*, 11–27. [[CrossRef](#)]
12. Khare, S.K.; Blanes-Vidal, V.; Nadimi, E.S.; Acharya, U.R. Emotion recognition and artificial intelligence: A systematic review (2014–2023) and research recommendations. *Inf. Fusion* **2024**, *102*, 102019. [[CrossRef](#)]
13. Hong, W.; Li, S.; Hu, Z.; Jiang, Q.; Weng, Y. Improving Relation Extraction by Knowledge Representation Learning. In Proceedings of the 2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI), Washington, DC, USA, 1–3 November 2021; pp. 1211–1215.
14. Kumar, A.; Jain, A.K. Emotion detection in psychological texts by fine-tuning BERT using emotion–cause pair extraction. *Int. J. Speech Technol.* **2022**, *25*, 727–743. [[CrossRef](#)]
15. Xu, X.; Yao, B.; Dong, Y.; Gabriel, S.; Yu, H.; Hendler, J.; Ghassemi, M.; Dey, A.K.; Wang, D. Mental-LLM: Leveraging Large Language Models for Mental Health Prediction via Online Text Data. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **2024**, *8*, 1–32. [[CrossRef](#)] [[PubMed](#)]
16. Nandwani, P.; Verma, R. A review on sentiment analysis and emotion detection from text. *Soc. Netw. Anal. Min.* **2021**, *11*, 81. [[CrossRef](#)] [[PubMed](#)]
17. Arshad, M.A.; Riaz, A.; Fatima, R.; Yasin, A. SevPredict: Exploring the Potential of Large Language Models in Software Maintenance. *AI* **2024**, *5*, 2739–2760. [[CrossRef](#)]
18. Charfaoui, K.; Mussard, S. Sentiment Analysis for Tourism Insights: A Machine Learning Approach. *Stats* **2024**, *7*, 1527–1539. [[CrossRef](#)]
19. Belbachir, F.; Roustan, T.; Soukane, A. Detecting Online Sexism: Integrating Sentiment Analysis with Contextual Language Models. *AI* **2024**, *5*, 2852–2863. [[CrossRef](#)]
20. Chung, H.W.; Hou, L.; Longpre, S.; Zoph, B.; Tay, Y.; Fedus, W.; Li, Y.; Wang, X.; Dehghani, M.; Brahma, S.; et al. Scaling instruction-finetuned language models. *J. Mach. Learn. Res.* **2024**, *25*, 1–53.

21. Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv* **2023**, arXiv:2307.09288.
22. Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *arXiv* **2025**, arXiv:2501.12948.
23. Liu, Y.; Han, T.; Ma, S.; Zhang, J.; Yang, Y.; Tian, J.; He, H.; Li, A.; He, M.; Liu, Z.; et al. Summary of chatgpt-related research and perspective towards the future of large language models. *Meta-Radiology* **2023**, *1*, 100017. [[CrossRef](#)]
24. Qin, X.; Wu, Z.; Zhang, T.; Li, Y.; Luan, J.; Wang, B.; Wang, L.; Cui, J. Bert-erc: Fine-tuning bert is enough for emotion recognition in conversation. In Proceedings of the AAAI Conference on Artificial Intelligence, Washington, DC, USA, 7–14 February 2023; Volume 37, pp. 13492–13500.
25. Luo, J.; Phan, H.; Reiss, J. Fine-tuned RoBERTa Model with a CNN-LSTM Network for Conversational Emotion Recognition. In Proceedings of the Interspeech 2023, Dublin, Ireland, 20–24 August 2023.
26. Douze, M.; Guzhva, A.; Deng, C.; Johnson, J.; Szilvassy, G.; Mazaré, P.E.; Lomeli, M.; Hosseini, L.; Jégou, H. The Faiss library. *arXiv* **2024**, arXiv:2401.08281.
27. Shapira, N.; Lazarus, G.; Goldberg, Y.; Gilboa-Schechtman, E.; Tuval-Mashiach, R.; Juravski, D.; Atzil-Slonim, D. Using computerized text analysis to examine associations between linguistic features and clients’ distress during psychotherapy. *J. Couns. Psychol.* **2021**, *68*, 77. [[CrossRef](#)]
28. Al Maruf, A.; Khanam, F.; Haque, M.M.; Jiyad, Z.M.; Mridha, F.; Aung, Z. Challenges and opportunities of text-based emotion detection: A survey. *IEEE Access* **2024**, *12*, 18416–18450. [[CrossRef](#)]
29. Hutto, C.; Gilbert, E. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In Proceedings of the International AAAI Conference on Web and Social Media, Ann Arbor, MI, USA, 1–4 June 2014; Volume 8, pp. 216–225.
30. Miller, G.A. WordNet: A lexical database for English. *Commun. ACM* **1995**, *38*, 39–41. [[CrossRef](#)]
31. Sebastiani, F.; Esuli, A. Sentiwordnet: A publicly available lexical resource for opinion mining. In Proceedings of the 5th International Conference on Language Resources and Evaluation. European Language Resources Association (ELRA), Genoa, Italy, 22–28 May 2006; pp. 417–422.
32. Hadi, M.U.; Qureshi, R.; Shah, A.; Irfan, M.; Zafar, A.; Shaikh, M.B.; Akhtar, N.; Wu, J.; Mirjalili, S.; Shah, M. A survey on large language models: Applications, challenges, limitations, and practical usage. *Authorea Prepr.* **2023**. [[CrossRef](#)]
33. Huang, Z.; Liang, M.; Qin, J.; Zhong, S.; Lin, L. Understanding self-attention mechanism via dynamical system perspective. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–3 October 2023; pp. 1412–1422.
34. Ghadekar, P.; Mohite, S.; More, O.; Patil, P.; Mangrulkar, S. Sentence Meaning Similarity Detector Using FAISS. In Proceedings of the 2023 7th International Conference on Computing, Communication, Control And Automation (ICCUBEA), Pune, India, 18–19 August 2023; IEEE: Scottsdale, AZ, USA, 2023; pp. 1–6.
35. Lima, F.F.d.; Osório, F.d.L. Empathy: Assessment instruments and psychometric quality—a systematic literature review with a meta-analysis of the past ten years. *Front. Psychol.* **2021**, *12*, 781346. [[CrossRef](#)] [[PubMed](#)]
36. Marchenko, O.; Radyvonenko, O.; Ignatova, T.; Titarchuk, P.; Zhelezniakov, D. Improving text generation through introducing coherence metrics. *Cybern. Syst. Anal.* **2020**, *56*, 13–21. [[CrossRef](#)]
37. Senbel, S. Fast and Memory-Efficient TFIDF Calculation for Text Analysis of Large Datasets. In *Advances and Trends in Artificial Intelligence. Artificial Intelligence Practices, Proceedings of the 34th International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2021*, Kuala Lumpur, Malaysia, 26–29 July 2021; Proceedings, Part I 34; Springer: Berlin/Heidelberg, Germany, 2021; pp. 557–563.
38. Villalobos, D.; Torres-Simón, L.; Pacios, J.; Paul, N.; Del Río, D. A systematic review of normative data for verbal fluency test in different languages. *Neuropsychol. Rev.* **2023**, *33*, 733–764. [[CrossRef](#)]
39. Rasool, A. Advancing the Intersection of AI and Bioinformatics. *J. Artif. Intell. Bioinform.* **2024**, *1*, 1–2. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.