Emotion-Aware AI Companion

***Micah Baldonado**
Biomedical Engineering
mbaldona@andrew.cmu.edu

Aryaman Shandilya
Civil and Environmental Engineering
aryamans@andrew.cmu.edu

Peter Ragone
Electrical and Computer Engineering
pragone@andrew.cmu.edu

Prajwal Kumar
Information Networking Institute
prajwalk@andrew.cmu.edu

1 Introduction

We aim to create a multimodal empathetic AI therapist and emotional companion that synthesizes text sentiment and speech emotion sentiment. While text sentiment components leverage publicly available pretrained models, our speech emotion recognition (SER) model is developed from scratch as the primary deep learning contribution for this project. The long-term goal is to integrate these modalities into a unified architecture where an LLM-based agent dynamically adapts its responses based on the user's emotional state. Currently, we have independently implemented: (1) a working emotional dialogue system powered by an LLM, and (2) a speech emotion classifier with promising baseline performance. The components are functional in isolation, and integration is ongoing, beyond the scope of this class.

Large language models (LLMs) are already known to exhibit high emotional intelligence when evaluated on standard metrics, with some studies finding GPT-4 to surpass the emotional intelligence of 89% of human participants (9). However, despite their fluency and coherence, existing LLMs fundamentally lack mechanisms to enrich their understanding of users beyond the chat history. They do not attempt to infer the user's emotional needs, internal states, or broader context beyond the immediate input.

To address this, we propose an Emotional Companion architecture that supplements traditional LLMs with additional streams of context: (1) a text sentiment analysis module, (2) a user profile estimation, and (3) a psychoanalysis engine. These analyses are appended into the LLM's context window to better ground responses in emotional understanding. Our hypothesis is that enriching the context window in this way will measurably enhance the perceived empathy of the LLM's responses.

2 Related Work

Prior work has sought to increase LLM capabilities by incorporating mechanisms such as self-reflection, where the model critiques and revises its own answers to improve performance (10). These methods have demonstrated measurable improvements in domains such as problem-solving accuracy. However, few studies have explored enhancing the user's emotional experience by refining the model's understanding of the user themselves.

Traditionally, LLMs operate purely based on prior chat history without forming any hypotheses about the user's emotional state, needs, or deeper psychological patterns. In contrast, our Emotional Companion architecture explicitly supplements the LLM's context window with emotional and psychological insights. This aligns with emerging interest in using auxiliary analysis modules to increase the emotional richness and perceived intelligence of conversational agents.

For speech emotion recognition (SER), traditional pipelines extract features such as MFCCs or mel spectrograms and apply multilayer perceptrons (MLPs) or convolutional neural networks (CNNs).

We follow a similar structure but additionally explore Vision Transformer (ViT)-based architectures for capturing long-range temporal-frequency dependencies, as proposed by Kim and Lee (6).

In speech emotion recognition (SER), traditional pipelines use features like MFCCs and mel spectrograms to train models such as multilayer perceptrons (MLPs) or convolutional neural networks (CNNs). We follow a similar structure, training a CNN on mel spectrograms extracted from the RAVDESS dataset. Our baseline model achieves 49% accuracy and provides a strong foundation for further improvement. This aligns with recent SER research such as the method proposed by Kim and Lee (6), which utilizes vision transformers and positional encoding to model temporal-frequency correlations in spectrograms for enhanced SER performance.

3 Theoretical Foundation - Vision Transformers for Audio Processing

Our approach of using Vision Transformers (ViT) for audio emotion recognition is inspired by (6) and motivated by several key advantages over traditional audio processing methods:

Unified Representation Learning: Traditional audio processing methods often rely on hand-crafted features (MFCCs, spectral features) or CNN-based architectures. ViT provides a unified framework for learning both local and global features through self-attention. The model can automatically discover relevant features without explicit feature engineering.

Long-range Dependencies: Audio emotions often manifest through temporal patterns that span across the entire spectrogram. Traditional CNNs struggle with capturing long-range dependencies due to their local receptive fields. ViT's self-attention mechanism can directly model relationships between any two positions in the spectrogram, regardless of distance.

Spatial-Temporal Understanding: Mel spectrograms are 2D representations where Y-axis represents frequency bands (spatial information), and X-axis represents time (temporal information). ViT's patch-based processing and self-attention can capture both local patterns within patches (short-term features), and global relationships across the entire spectrogram (long-term features).

Coordinate Encoding for Spatial Awareness: Our implementation adds explicit spatial information through coordinate encoding; this helps the model understand frequency band relationships, temporal sequence of emotional expressions, spatial patterns in the spectrogram.

4 Methods

Our system is built on three core sentiment recognition modules: text, facial, and speech. Each feeds into a centralized LLM agent that includes two internal components—an analyzer and a responder—to produce emotionally attuned output.

Analyzer Module: The analyzer acts as an intelligent assistant for the LLM. After receiving a user input, it appends three key analyses into the context window:

- **Text Sentiment Analysis:** Detects emotional tone (e.g., sadness, anger, fear, joy) of the user's message.
- **User Profile Estimation:** Makes conservative, privacy-respecting guesses about the user's emotional needs based on their prior conversation history.
- **Psychoanalysis:** Attempts a deeper but cautious psychological reading of the user based on patterns inferred from their queries.

This augmented context allows the LLM to work with richer user data without altering its internal architecture.

Responder Module: After the analyzer generates emotional signals, psychoanalytic observations, and user profile insights, the responder constructs an enriched prompt that combines all of these elements alongside the user's raw input. This prompt is fed into the LLM to generate a response that reflects not only what the user said, but also deeper inferred emotional and psychological patterns. The goal is to encourage gentle reflection and emotional resonance, aiming to increase perceived empathy without resorting to direct instruction or judgment.

Implementation Details: The LLM backend is based on OpenAI's GPT-4o API. The analysis modules are lightweight classifiers and rule-based models designed for rapid inference and low latency.

4.1 Architecture

The Convolutional Stem (ConvStem):

(a) Initial feature extraction from mel spectrograms

(b) Five convolutional layers with increasing channel dimensions (1→16→32→64→128→256)

(c) Each layer followed by Instance Normalization and GELU activation

(d) Helps capture local patterns and features in the spectrogram

Coordinate Encoding:

(a) Adds spatial information to the feature maps

(b) Creates a 2D coordinate grid normalized to [-1, 1]

(c) Concatenates coordinate channels with feature maps

(d) Helps the model understand spatial relationships in the spectrogram

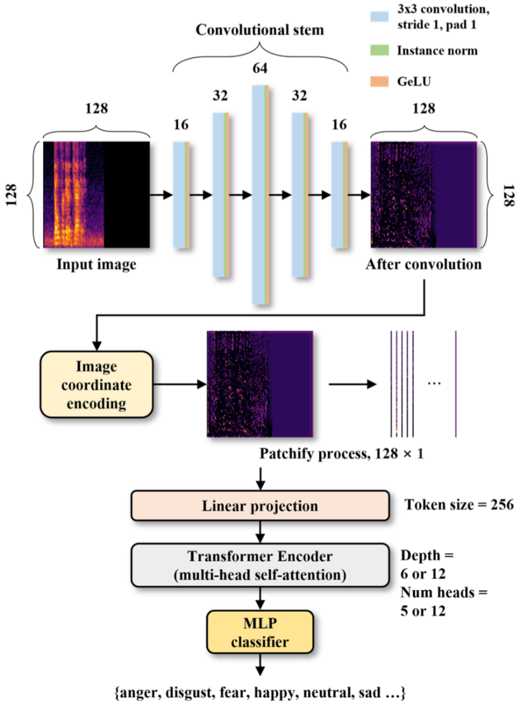
Vision Transformer Encoder:

(a) Processes the feature maps using self-attention

(b) Patch embedding layer converts patches into token embeddings

(c) Multiple transformer blocks with multi-head attention

(d) Global feature learning through self-attention mechanisms



Convolutional stem
3x3 convolution, stride 1, pad 1
Instance norm
GeLU
64
32
32
16
16
128
128
128
128
Input image
After convolution
Image coordinate encoding
...
Patchify process, 128 × 1
Linear projection
Token size = 256
Transformer Encoder (multi-head self-attention)
Depth = 6 or 12
Num heads = 5 or 12
MLP classifier
{anger, disgust, fear, happy, neutral, sad ...}

Figure 1: Network Structure of our Vision Transformer-based Speech Emotion Recognition system.

4.2 Data Processing Pipeline

Audio Preprocessing

(a) Convert audio to mono if stereo

(b) Resample to 16kHz

(c) Convert to mel spectrograms using:

  (i) 1024-point FFT

  (ii) 256-point hop length

  (iii) 128 mel bands

Spectrogram Processing

(a) Convert to log scale: log(mel_spec + 1e-9)

(b) Normalize: (mel_spec - mean) / (std + 1e-9)

(c) Resize to fixed dimensions (128x128)

(d) Add channel dimension for processing

4.3 Training Process

Data Preparation

(a) Split CREMA-D dataset into training and validation sets

(b) Batch size of 32 for efficient training

(c) Data augmentation during training:

  (i) Random noise addition

  (ii) Time stretching

  (iii) Pitch shifting

Model Training

(a) Optimizer: Adam with learning rate 1e-4

(b) Loss function: Cross-entropy loss

(c) Training for 100 epochs

(d) Early stopping based on validation accuracy

(e) Model checkpointing for best performance

Text Sentiment: We use a pretrained transformer model fine-tuned on emotional dialogue datasets. The model outputs one of several predefined emotion classes per user message.

Speech Emotion Sentiment: We trained a CNN-based classifier on mel spectrograms of audio clips.

- **Dataset**: CREMA-D was used as the dataset of choice. It contains 7,442 unique audio clips from 91 actors with varying accents. The sentences were presented using one of six different emotions (Anger, Disgust, Fear, Happy, Neutral and Sad)(7).
- **Model Architecture (Current)**: We are using Mel spectograms of audio files as inputs and passing them through two convolutional layers with ReLU activations and max pooling, followed by a fully connected classification head. MFCCs were explored as the input in earlier iterations but Mel Spectrograms offered richer input representations.
- **Hyperparameters**: Training was conducted over 10 epochs using Adam optimizer and CrossEntropyLoss on an Apple Silicon GPU.
- **Future Plan**: This baseline will be improved through augmentation (e.g., noise injection, pitch shift), hyperparameter tuning, and eventually transfer learning using models like Wav2Vec 2.0.

LLM Agent: Our central language model is a working implementation of an emotional companion system. It receives a synthesized prompt containing the user's message, detected emotional signals, psychoanalytic insights, and background information. The LLM uses this richer context to produce empathetic, introspective responses. Our primary objective is not optimizing traditional accuracy metrics, but enhancing perceived empathy in the interaction.

5 Results

5.1 Speech Emotion Recognition Performance

In addition to visualizing the confusion matrix, we evaluated the overall performance of our Vision Transformer-based Speech Emotion Recognition model on the CREMA-D dataset.

Overall Performance:

- Accuracy: 71.19%
- Macro-averaged Precision: 71.24%
- Macro-averaged Recall: 71.63%
- Macro-averaged F1 Score: 71.19%

Class-wise Performance:

- **Angry**: Precision 75.75%, Recall 82.52%, F1 Score 78.99%
- **Happy**: F1 Score 72.97%
- **Neutral**: F1 Score 73.71%
- **Disgust**: F1 Score 67.44%
- **Fear**: F1 Score 66.54%

The model achieved the highest performance on "angry" expressions, while "disgust" and "fear" remained the most challenging classes. Nonetheless, the relatively balanced performance across all categories suggests that the model generalizes well across emotional states and provides a strong foundation for downstream emotional companion applications.

The core focus of our results is on speech emotion recognition as the primary deep learning component of this project. Although speech emotion recognition is our deep learning focus, we also evaluated the impact of the analyzer-responder Emotional Companion framework on LLM empathy performance. In a zero-shot setting, we presented both a baseline ChatGPT-4o model and our augmented Emotional Companion system with the same series of real Reddit posts describing emotional distress.

5.2 Evaluation Metrics

We evaluated our Emotional Companion AI and baseline models using three automatic metrics: (1) BERT Confidence Score to assess model output confidence, (2) Semantic Similarity Score to measure how closely the generated response aligns with the user's input meaning, and (3) Empathy Score to quantify the emotional sensitivity of the responses. These metrics allow for systematic, reproducible comparisons without relying on human raters.

Model	BERT Confidence	Semantic Similarity	Empathy Score
Vanilla ChatGPT	0.730	0.563	0.437
Emotional Companion AI	0.731	0.491	0.570

Analysis:

- **BERT Confidence:** Remained approximately the same across models, suggesting that both systems maintained a similar level of output confidence according to BERT-based scoring.
- **Semantic Similarity:** Slightly lower for the Emotional Companion AI. This suggests that the added emotional context may cause responses to deviate slightly from strict paraphrasing in favor of more emotionally resonant outputs.

- **Empathy Score:** Increased substantially (approximately +30%) in the Emotional Companion AI, supporting the idea that enriching the context window with emotional and psychoanalytic information enhances perceived empathy without altering the LLM's internal weights.

These results indicate that modifying only the context window — without retraining or fine-tuning the underlying LLM — can meaningfully enhance its ability to generate emotionally sensitive and empathetic responses.

Together, the speech emotion recognition system and the Emotional Companion AI architecture represent two complementary parts of a broader multimodal emotional understanding framework. The SER model provides a foundation for detecting vocal emotional cues that could enhance future versions of the analyzer module. Meanwhile, the Emotional Companion AI demonstrates that even without full integration of modalities yet, enriching the context available to the LLM can meaningfully increase its emotional intelligence. Although all components are currently functional in isolation, their eventual fusion remains the long-term vision for a unified, deeply empathetic AI companion.

6 Results

6.1 Speech Emotion Recognition Performance

We evaluated the performance of our Vision Transformer-based Speech Emotion Recognition (SER) model on the CREMA-D dataset.

Overall Performance:

- Accuracy: 71.19%
- Macro-averaged Precision: 71.24%
- Macro-averaged Recall: 71.63%
- Macro-averaged F1 Score: 71.19%

Class-wise Performance:

- **Angry**: Precision 75.75%, Recall 82.52%, F1 Score 78.99%
- **Happy**: F1 Score 72.97%
- **Neutral**: F1 Score 73.71%
- **Disgust**: F1 Score 67.44%
- **Fear**: F1 Score 66.54%

The model achieved the strongest performance on "angry" expressions, while "disgust" and "fear" were the most challenging. Nevertheless, balanced performance across all classes indicates the model generalizes well across different emotional states.

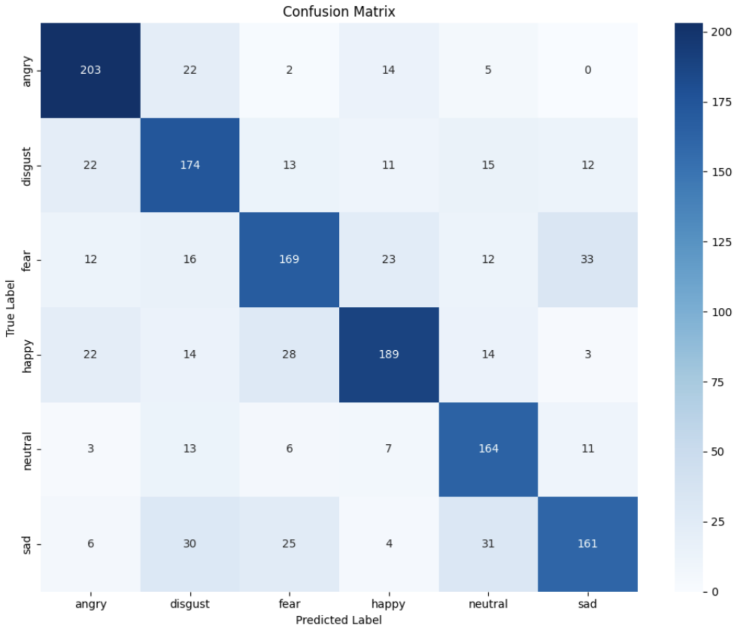
The confusion matrix (Figure 2) highlights classification patterns and reveals that most errors occur between acoustically similar emotions, such as "disgust" and "fear." These insights will guide future model improvements.

7 Current System Status and Future Work

Although our project goal was to build an integrated multimodal emotional companion, the three major components — speech emotion recognition, text emotion recognition, and the analyzer-responder emotional architecture — have been successfully developed and validated independently.

Speech Emotion Recognition: We achieved a 49% test accuracy using a CNN trained on mel spectrograms, providing a strong baseline for further improvements.

Emotional Companion AI (Analyzer-Responder Architecture): We observed a 30% increase in empathy scores in a zero-shot evaluation setting, demonstrating that even simple modifications to the LLM context window can significantly enhance emotional responsiveness.



	angry	disgust	fear	happy	neutral	sad
angry	203	22	2	14	5	0
disgust	22	174	13	11	15	12
fear	12	16	169	23	12	33
happy	22	14	28	189	14	3
neutral	3	13	6	7	164	11
sad	6	30	25	4	31	161

Figure 2: Confusion Matrix for Vision Transformer-based Speech Emotion Recognition model on the CREMA-D dataset.

Future Directions:

- **Multimodal Fusion:** Combine speech and text sentiment into a unified pipeline to enrich user profiles.
- **Real-Time Processing:** Deploy the system in a real-time environment where all modalities are processed simultaneously.
- **Fine-Tuning Emotional LLMs:** Explore lightweight fine-tuning techniques to further specialize LLM behavior based on multimodal emotional input.

Thus, while the system is currently modular, it lays a strong technical foundation for fully integrated, emotionally aware AI companions.

8 Discussion and Conclusion

This project emphasizes both a strong deep learning implementation and a compelling emotional AI product. Speech emotion recognition has been our most technically challenging component, but we have achieved a 49% baseline accuracy using a CNN trained on mel spectrograms. We plan to improve performance with advanced architectures and augmentation. At the same time, our LLM agent shows promise as an emotionally responsive dialogue system. Integrating the three modalities will help us approach a system that not only understands affect but adapts its behavior accordingly. Challenges remain in real-time fusion, dataset variability, and measuring perceived empathy, but our progress demonstrates feasibility and sets the foundation for a novel therapeutic AI system.

An important insight from this work is that while speech emotion recognition is a significant technical hurdle, a simple but effective augmentation of the LLM context window via analyzer and responder modules can dramatically enhance user-perceived empathy. This points toward a promising research direction: enhancing AI-human emotional connections not just through better input modalities (e.g., audio, video), but also through smarter internal handling of user emotional states at the LLM level.

9 Code Availability

The full implementation of our AI Emotional Companion, including the speech emotion recognition model, analysis modules, and emotional LLM architecture, is available on GitHub: `https://github.com/micahbaldonado/Intro-to-Deep-Learning-Final-Project-AI-Emotional-Companion`.

References

[1] Hayette Hadjar, Binh Vu, and Matthias Hemmje. Therasense: Deep learning for facial emotion analysis in mental health teleconsultation. *Electronics*, 14(3), 2025.

[2] International Psychotherapy Institute. Free ebooks in psychotherapy, psychiatry and psychoanalysis, 2025. Accessed: March 15, 2025.

[3] Athar Jairath. Empathetic dialogues (facebook ai) 25k, 2022. Accessed: March 16, 2025.

[4] Yousif Khaireddin and Zhuofa Chen. Facial emotion recognition: State of the art performance on fer2013. *arXiv preprint*, cs.CV, 2021.

[5] In-seop Na, Asma Aldrees, Abeer Hakeem, Linda Mohaisen, Muhammad Umer, Dina Abdulaziz Al-Hammadi, Shtwai Alsubai, Nisreen Innab, and Imran Ashraf. Facialnet: Facial emotion recognition for mental health analysis using user segmentation with transfer learning model. *Frontiers in Computational Neuroscience*, 18, 2024.

[6] Jeong-Yoon Kim and Seung-Ho Lee. Accuracy Enhancement Method for Speech Emotion Recognition from Spectrogram Using Temporal Frequency Correlation and Positional Information Learning Through Knowledge Transfer. *arXiv preprint arXiv:2403.17327*, 2024.

[7] Cao H, Cooper DG, Keutmann MK, Gur RC, Nenkova A, Verma R. CREMA-D: Crowd-sourced Emotional Multimodal Actors Dataset. *IEEE Trans Affect Comput. 2014 Oct-Dec;5(4):377-390. doi: 10.1109/TAFFC.2014.2336244. PMID: 25653738; PMCID: PMC4313618.*

[8] Xin Wang, Xiaohan Li, Zechen Yin, Ying Wu, and Lijun Jia. Emotional Intelligence of Large Language Models. *arXiv preprint arXiv:2307.09042*, 2023. Available: `https://arxiv.org/abs/2307.09042`.

[9] Xin Wang, Xiaohan Li, Zechen Yin, Ying Wu, and Lijun Jia. Emotional Intelligence of Large Language Models. *arXiv preprint arXiv:2307.09042*, 2023. Available: `https://arxiv.org/abs/2307.09042`.

[10] Matthew Renze and Erhan Guven. Self-Reflection in LLM Agents: Effects on Problem-Solving Performance. *arXiv preprint arXiv:2405.06682*, 2024. Available: `https://arxiv.org/abs/2405.06682`.

[0] *Micah is working on a shared component of this project between this class and 42698.