

Graphical Abstract

Decoding Neural Emotion Patterns through Natural Language Processing Embeddings

Gideon Vos, Maryam Ebrahimpour, Liza van Eijk, Zoltan Sarnyai, Mostafa Rahimi Azghadi

Decoding Neural Emotion Patterns through Natural Language Processing Embeddings

1



This study introduces a computational framework for directly mapping natural language emotional content to brain regions without requiring neuroimaging.

2



The framework demonstrated high spatial specificity by accurately mapping 27 discrete emotions to neuro-anatomically plausible brain regions.

3



Regional assignment patterns from textual data showed strong consistency with established neuroimaging research.

4



The integration of semantic embeddings and neuro-anatomical mapping successfully differentiated between healthy and depressed populations through distinct limbic activation patterns.

Highlights

Decoding Neural Emotion Patterns through Natural Language Processing Embeddings

Gideon Vos, Maryam Ebrahimpour, Liza van Eijk, Zoltan Sarnyai, Mostafa Rahimi Azghadi

- This study introduces a computational framework for directly mapping natural language emotional content to brain regions without requiring neuroimaging.
- The integration of semantic embeddings and neuro-anatomical mapping successfully differentiated between healthy and depressed populations through distinct limbic activation patterns.
- The framework demonstrated high spatial specificity by accurately mapping twenty-seven discrete emotions to neuro-anatomically plausible brain regions.
- Regional assignment patterns showed strong consistency with established neuroimaging research.
- In favor of reproducible research and to advance the field, all programming code used in this study is made publicly available.

Decoding Neural Emotion Patterns through Natural Language Processing Embeddings

Gideon Vos^a, Maryam Ebrahimpour^a, Liza van Eijk^b, Zoltan Sarnyai^c,
Mostafa Rahimi Azghadi^a

^a*College of Science and Engineering, James Cook University, James Cook
Dr, Townsville, 4811, QLD, Australia*

^b*College of Health Care Sciences, James Cook University, James Cook
Dr, Townsville, 4811, QLD, Australia*

^c*College of Public Health, Medical, and Vet Sciences, James Cook University, James
Cook Dr, Townsville, 4811, QLD, Australia*

Abstract

Introduction. Understanding the neural correlates of emotional expression in natural language represents a significant challenge in computational neuroscience and affective computing. While traditional neuroimaging studies require expensive equipment and controlled laboratory environments, the increasing availability of digital text data presents new opportunities for emotion-brain mapping. Previous research has primarily focused on either neuroimaging-based emotion localization or computational text analysis independently, with limited integration between these domains. This study proposes a novel computational approach that, while not validated against direct neuroimaging, explores potential relationships between textual emotional content and anatomically-defined brain regions.

Methods. We developed a computational pipeline that maps textual emotion expressions to specific brain regions through a multi-stage process. The framework utilizes OpenAI’s text-embedding-ada-002 model to generate high-dimensional semantic representations of input texts, followed by dimensionality reduction and clustering to identify emotional clusters. These clusters are then mapped to 18 predefined neuro-anatomic brain regions associated with emotional processing. Three distinct experiments were conducted. First, we analyzed conversational data from healthy and depressed subjects using the Distress Analysis Interview Corpus/Wizard-of-Oz (DIAC-WOZ) dataset, comparing emotional brain mapping patterns between these populations.

Next, we repeated this process the comprehensive GoEmotions classification dataset. Emotional intensity was quantified using a lexical scoring system that evaluates keyword presence, syntactic patterns, and linguistic modifiers. Finally, we performed a comparison between human-produced text and the response generated by a Large Language Model (LLM) chat bot to evaluate differences in emotional brain activation patterns and determine the extent to which AI-generated language mirrors human emotional expression.

Results. Our proposed approach successfully mapped textual emotions to neuro-anatomically plausible brain regions with high spatial specificity across experiments. Distinct activation patterns emerged between healthy and depressed populations, with depressed subjects showing increased engagement of limbic regions associated with negative affect processing. Extended emotion analysis demonstrated successful differentiation of discrete emotional states. Our final experimental results revealed that while LLM-generated responses exhibited a similar distribution of basic emotions, they lacked the nuanced regional activation observed in human text, particularly in areas associated with empathy and self-referential processing such as the medial prefrontal cortex and posterior cingulate cortex.

Conclusion. This study presents a computational framework for directly mapping natural language emotional content to brain regions without requiring neuroimaging data. The novel integration of semantic embeddings, unsupervised clustering, and neuro-anatomical mapping provides a scalable approach for emotion-brain research that can process large-scale textual datasets. The framework’s ability to distinguish between healthy and depressed populations, as well as differentiate among discrete emotions, demonstrates its potential for both clinical applications and basic emotion research. The methodology offers significant advantages over traditional neuroimaging approaches, including cost-effectiveness, scalability, and the ability to analyze naturalistic language data. Finally, this framework provides a brain-inspired benchmark for evaluating how closely AI-generated language mirrors human emotional expression by comparing their inferred neural activation patterns.

Keywords: Artificial Intelligence, Mental Health, Depression

PACS: 07.05.Mh, 87.19.La

2000 MSC: 68T01, 92-08

1. Introduction

Understanding the neural correlates of emotion has traditionally relied on neuroimaging modalities such as electroencephalography (EEG) and functional magnetic resonance imaging (fMRI) [1–3]. These techniques have revealed the involvement of regions like the amygdala, insula, anterior cingulate cortex, and prefrontal cortex across different emotional states [4]. Meta-analyses of neuroimaging studies have consistently identified these key regions across diverse emotional paradigms, with the amygdala showing particular importance for threat detection and fear processing [5], while the anterior cingulate cortex and insula demonstrate critical roles in emotional awareness and interoceptive processing [6, 7]. However, traditional neuroimaging approaches face significant limitations including high costs, restricted accessibility, controlled laboratory requirements, and limited ecological validity when studying naturalistic emotional expression [8, 9]. The increasing availability of digital text data presents unprecedented opportunities for emotion-brain mapping that could overcome these existing constraints.

Caucheteux *et al.* [10] demonstrated that large language model (LLM) embeddings align with human brain activity without fine-tuning, showing that pre-trained language models inherently capture aspects of neural language representations. This foundational work was extended by Toneva *et al.* [11], who introduced the concept of brain embeddings, highlighting the geometric alignment between the representational spaces of LLMs and brain activity during reading and listening tasks.

Further evidence of this alignment comes from Schrimpf *et al.* [12], who showed that artificial neural networks, especially transformer-based architectures as used in LLMs, can predict human neural responses to language with remarkable accuracy. These findings collectively support the feasibility of leveraging LLM-derived embeddings to model brain activity, suggesting that the semantic representations learned by these models may reflect fundamental aspects of how the human brain processes language and emotion.

Parallel efforts have explored mapping emotional text representations to neuro-biological substrates. Tomasino *et al.* [13] developed a cognitive-affective framework demonstrating how emotionally charged linguistic input recruits distinct neural systems, while Chen *et al.* [14] correlated sentiment

analysis outputs with fMRI patterns, confirming the brain’s differentiation of emotional valence during narrative comprehension. These studies build upon earlier work demonstrating that emotional processing involves distributed neural networks, with positive emotions preferentially engaging left prefrontal regions and negative emotions showing stronger right hemisphere activation [15].

Zhou *et al.* [16] extended this work by associating semantic embeddings from emotional narratives with fMRI-derived brain states, particularly noting strong alignment in medial prefrontal and temporal regions. Similarly, Xiao *et al.* [17] applied unsupervised clustering on emotion-labeled text to uncover latent emotional dimensions and correlated them with EEG and fMRI features. These computational approaches align with neuroimaging meta-analyses showing that different emotional categories activate distinct but overlapping brain networks, with cognitive emotions recruiting prefrontal cortical areas [18, 19].

The utility of natural text has been further emphasized by Hoemann *et al.* [20], who highlighted the value of social media and dialogue data for studying emotion in real-world contexts. Their findings support using large-scale spontaneous language datasets to infer affective brain states, moving beyond the artificial constraints of laboratory-based emotion elicitation paradigms. This approach is particularly relevant given further research showing that naturalistic emotional expression differs significantly from laboratory-induced emotions in both linguistic patterns and associated neural activity [20].

Despite these advances, existing studies have not established a fully computational, imaging-free method that directly links textual emotional content to specific neuro-anatomical regions. Current approaches typically require either controlled laboratory settings that limit validity, or a focus on general emotional dimensions rather than specific brain region mapping [8].

This represents a significant gap in our ability to study emotional processing at scale. The motivation for developing a purely computational emotion-brain mapping framework therefor stems from several converging factors. First, the exponential growth of digital text data offers an unprecedented window into human emotional expression in naturalistic contexts [20]. Second, advances in semantic embedding technologies have created powerful

tools for capturing nuanced relationships between language, meaning, and affective content. Third, decades of neuroimaging research have established a robust understanding of the neuro-anatomical basis of emotion processing, with meta-analytic evidence highlighting key neural circuits involved in emotional regulation and expression [18, 21].

The theoretical foundation of our proposed approach rests on the principle that emotional expression in language reflects underlying neural processes. This hypothesis is supported by evidence showing that the human brain encodes language through distributed, continuous representations [22–24]. Recent studies demonstrated a direct alignment between high-dimensional language model embeddings used by LLMs and neural activation patterns in language-processing regions [23–25]. Such findings suggest that embedding spaces learned by LLMs may mirror the brain’s own semantic encoding mechanisms, indirectly capturing semantic and emotional features that align with patterns of neural activity.

Furthermore, individuals with certain mental health conditions may exhibit distinct language patterns, including variations in word choice, emotional tone, and syntactic complexity that correlate with well-documented neurobiological abnormalities [8, 26–31]. Extending this perspective, characteristic language patterns in clinical populations may manifest as distinct clusters within embedding space, reflecting altered cognitive and affective processing [30, 32–34]). If these deviations also correspond to measurable shifts in brain activation, embedding-based models could offer novel insights into brain-language relationships and help inform computational mental health diagnostics.

Beyond clinical contexts, this framework could further support advanced applications in areas such as LLM-generated text detection [35], where subtle differences in linguistic patterns, coherence, and semantic clustering could indicate non-human authorship.

This study therefor investigates whether it is possible to computationally map natural language emotional expressions directly to brain regions without neuroimaging data, using semantic embeddings and clustering techniques to bridge text analysis with neuro-anatomical mapping. Specifically, we aim to:

- Develop a novel computational framework that transforms textual emotional content into neuro-anatomically plausible brain region activations using state-of-the-art natural language processing techniques.
- Validate the clinical utility of this approach by demonstrating its ability to differentiate between healthy and depressed populations through distinct emotional brain mapping patterns, building on established neuroimaging differences in depression.
- Demonstrate discrete emotion localization by mapping specific emotional states to neuro-anatomically appropriate brain regions with high spatial specificity, consistent with established emotion-brain mapping literature.
- Provide an objective, brain-based measure of human-likeness in language by comparing AI-generated and human-authored texts through their inferred emotional brain activation patterns.

By providing a cost-effective, scalable alternative to traditional neuroimaging methods, such a computational approach can open new avenues for understanding the neural basis of human emotional communication in digital environments, while maintaining the ability to generate interpretable, neuro-anatomically grounded predictions from textual data alone.

2. Methods

2.1. Datasets, Preprocessing and Text Preparation

Three text-based datasets were employed in this study (Table 1). The DIAC-WOZ dataset [36] comprises annotated interview transcripts from individuals diagnosed with depression and healthy controls. The GoEmotions dataset [37] includes 58,000 Reddit [38] comments manually labeled into 27 emotion categories (or neutral). The Schema-Guided Dialogue dataset [39] represents nearly half a million sentences comprised of human and LLM chat bot interactions. All datasets consist of texts produced by native English speakers.

Table 1: Datasets utilized in this study.

Dataset	Emotions	Subjects
DAICWOZ[36]	Healthy and Depressed Categories	134 Clinical interview transcripts
GoEmotions [37]	27 Emotion Categories	58k English Reddit comments
The Schema-Guided Dialogue Dataset [39]	Human and Chat bot conversations	463,282 English sentences

2.2. Text Embedding Generation

During preprocessing (Figure 1, step 1), text was segmented into chunks of approximately 300 characters using periods as sentence boundaries. Next, text embeddings were generated using OpenAIs [40] *text-embedding-ada-002* model (Figure 1, step 2), which produces 1536-dimensional vector representations of input text. This model was chosen for its strong performance in capturing both semantic relationships and emotional nuances in natural language [28–31]. Although training a custom embeddings model for our experimentation is technically feasible, it could introduce unwanted bias into our experiments, while the OpenAI *text-embedding-ada-002* embeddings are commonly used in many public and commercial LLMs.

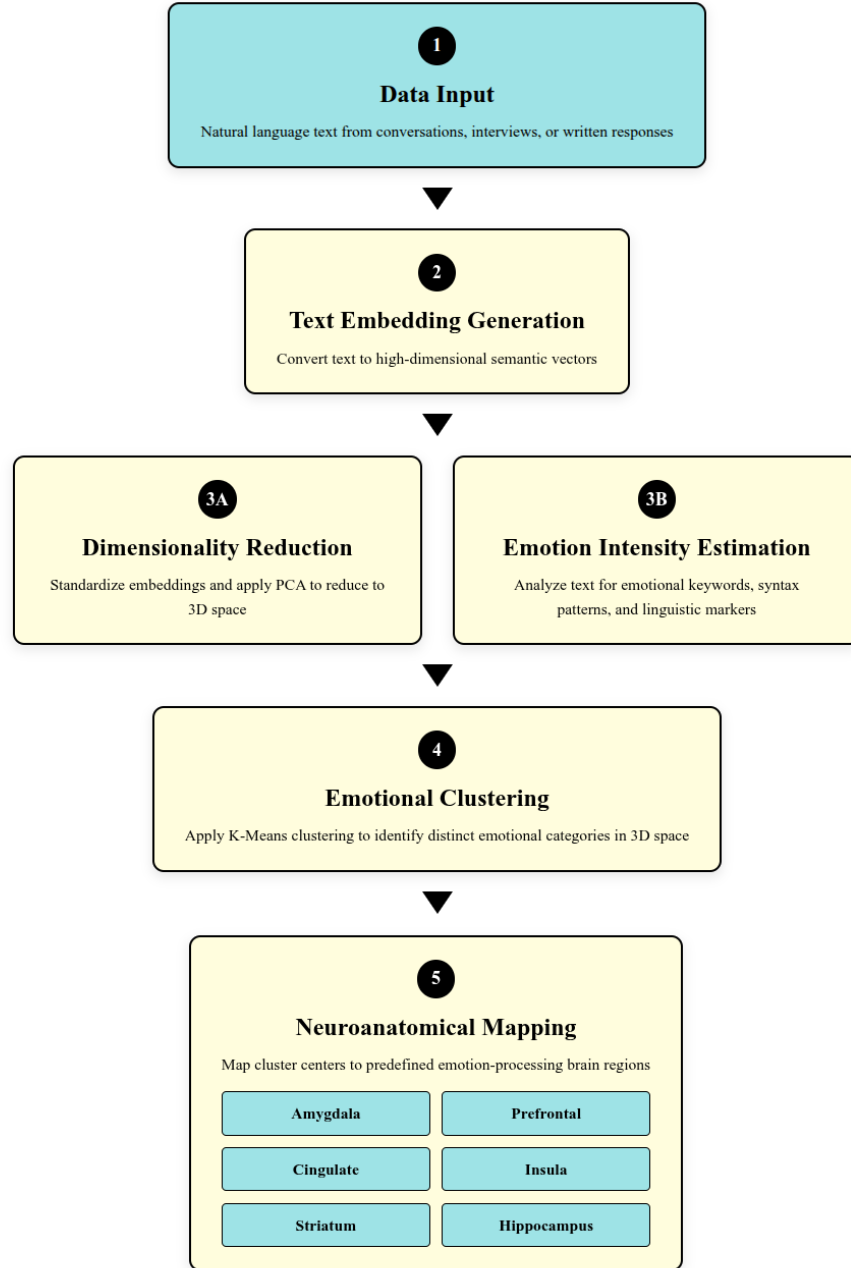


Figure 1: Five-step text computational pipeline to convert natural language text to embeddings, reduce dimensionality, cluster to emotional groups and map to final brain regions.

2.3. Dimensionality Reduction and Spatial Mapping

The high-dimensional embeddings underwent a dimensionality reduction process (Figure 1, step 3A) using Principal Component Analysis (PCA) to reduce the dimensionality to three components, representing the minimum number of dimensions required for spatial brain mapping.

2.4. Emotional Intensity Estimation

Emotional intensity was quantified using a lexicon-based approach [41–46] combined with syntactic feature analysis (Figure 1, step 3B). This emotion lexicon was constructed containing multiple terms across intensity levels:

- Extreme intensity (1.0): Terms indicating maximum emotional activation (e.g., "devastated," "euphoric," "depressed," "extremely")
- High intensity (0.8): Strong emotional indicators (e.g., "amazing," "hate," "terrible")
- Moderate intensity (0.6): Common emotional expressions (e.g., "love," "sad," "happy")
- Mild intensity (0.3): Subtle emotional content (e.g., "nice," "bad," "okay")

Additional intensity modifiers were incorporated to capture amplification effects:

- Intensification modifiers ("so," "very," "really," "truly," "completely," "totally") added 0.3 to base intensity
- Absolutist terms ("never," "always," "everything," "nothing") contributed 0.2 additional intensity
- Exclamation marks added 0.25 per occurrence (maximum 4)
- Question marks contributed 0.15 per occurrence (maximum 3)
- All-caps text (more than 3 characters) added 0.5 intensity

Base intensity was set at 0.1 for all texts to account for implicit emotional content, with final intensity scores capped at 2.0 to prevent extreme outliers from skewing subsequent analyses.

2.5. Emotion Region Clustering

K-means clustering was applied to the 3D PCA-transformed embeddings to identify distinct emotional patterns within the data (Figure 1, step 4). The number of clusters was set to match the number of predefined anatomical brain regions (11 bilateral + 7 midline = 18 regions), establishing a direct correspondence between emotional content clusters and neuro-anatomical structures [1, 5, 47–49]. Of the 18 anatomically defined brain regions selected, 14 have been consistently implicated in emotion processing [1–3, 50].

2.6. Cluster-to-Region Assignment

The assignment of emotional content clusters to specific brain regions (Figure 1, step 5) employed a two stage distance minimization approach that preserved the one-to-one mapping between clusters and anatomical regions.

First, each cluster center’s coordinates in the 3D PCA space were compared against the predefined Montreal Neurological Institute (MNI) coordinate positions of the 18 emotion processing brain regions using Euclidean distance. These regions were selected to span key nodes of the emotional circuitry and included limbic structures, prefrontal regions, subcortical structures, temporal regions, and brain stem nuclei, along with anterior and posterior cingulate cortex and medial prefrontal cortex.

The mapping process utilized a greedy assignment algorithm where cluster centers were sequentially matched to their nearest available brain regions. For each cluster center, distances to all anatomical regions were calculated and sorted in ascending order. The cluster was then assigned to the closest region that had not yet been claimed by another cluster, with this constraint preventing multiple clusters from mapping to the same anatomical location.

This sequential assignment continued until all clusters were mapped to unique brain regions, ensuring that the spatial distribution of emotional content in the 3D space corresponded meaningfully to the anatomical organization of emotion-processing brain networks. Once cluster-to-region assignments were established, individual text samples inherited their brain region labels based on their cluster membership, creating the final mapping from textual content to neuro-anatomical locations.

2.7. Statistical Analysis

The computational pipeline incorporated statistical practices including random seed setting to ensure reproducibility and management of edge cases such as insufficient sample sizes. Region-specific analysis was conducted by aggregating texts assigned to each brain region and calculating mean emotional intensities, providing quantitative measures of regional emotional activation patterns. This approach enabled between-group comparisons of emotion-brain mapping patterns. Figure 2 provides a visual hierarchy detailing the emotion to brain region mapping approach as detailed in steps 4 and 5 of Figure 1.

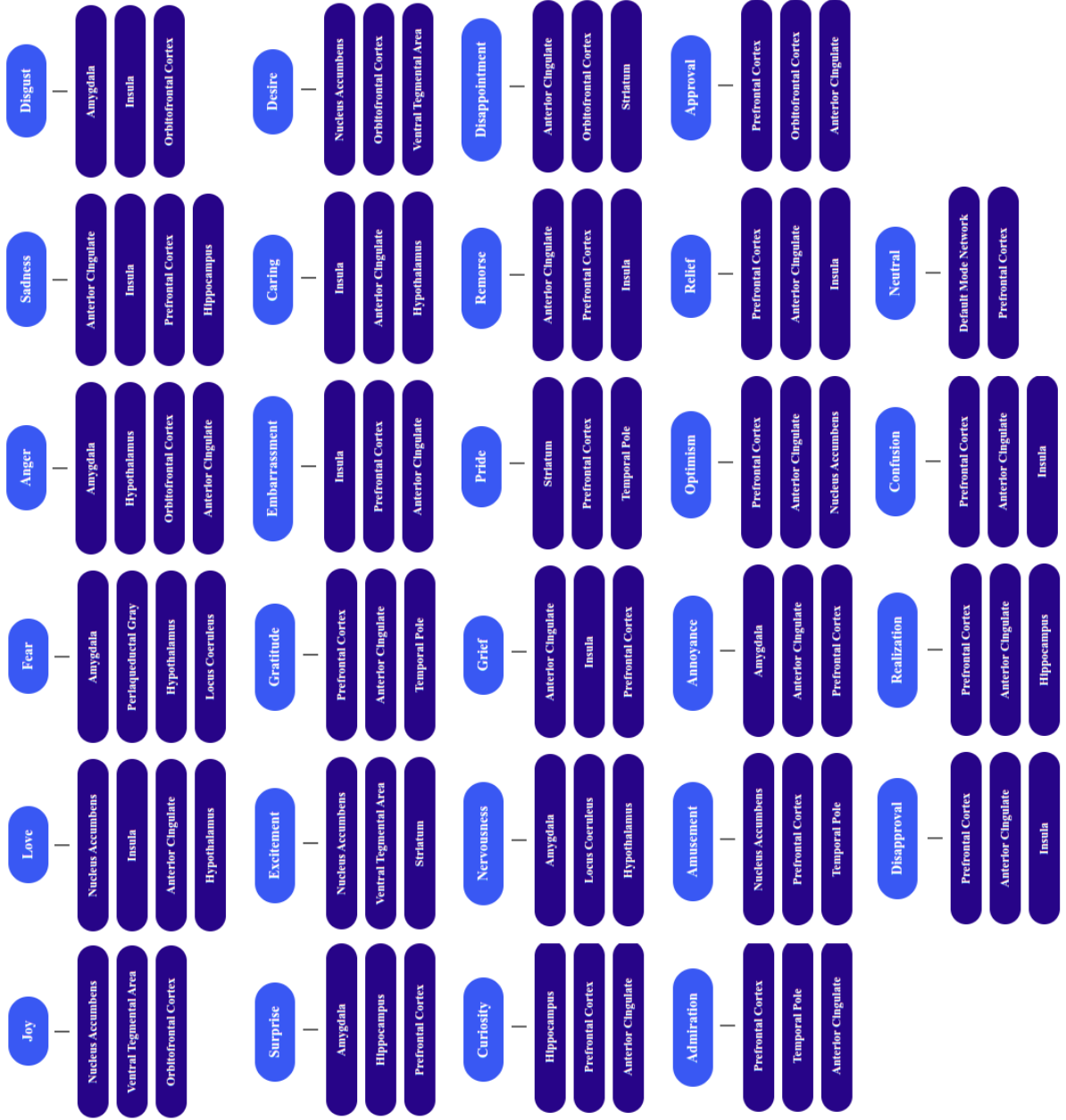


Figure 2: Emotion to brain region assignment hierarchy applied in this study.

The programming pseudo-code for the three key algorithms utilized and described in Section 2 are detailed below:

Algorithm 1 Text Preprocessing and Embedding Generation

Input: Text datasets $D = \{d_1, d_2, d_3\}$

Output: 1536-dimensional embeddings matrix

```

1: // Step 1: Text Preprocessing and Chunking
2: function PREPROCESSTEXTS(texts)
3:   chunks  $\leftarrow$  []
4:   for each text in texts do
5:     segments  $\leftarrow$  split text into  $\approx 300$  character chunks using periods
6:     chunks.append(segments)
7:   end for
8:   return chunks
9: end function
10:
11: // Step 2: Text Embedding Generation
12: function GETADAEMBEDDINGS(texts)
13:   Initialize OpenAI client with API key
14:   embeddings  $\leftarrow$  [], batch_size  $\leftarrow$  2000
15:   for  $i = 0$  to  $\text{len}(\text{texts})$  step batch_size do
16:     batch  $\leftarrow$  texts[ $i : i + \text{batch\_size}$ ]
17:     response  $\leftarrow$  client.embeddings.create(model="text-embedding-ada-002", input=batch)
18:     batch_embeddings  $\leftarrow$  extract embeddings from response
19:     embeddings.extend(batch_embeddings)
20:   end for
21:   return np.array(embeddings) // Shape: (n_samples, 1536)
22: end function

```

Algorithm 2 Dimensionality Reduction and Emotional Intensity Estimation

Input: High-dimensional embeddings from Step 2.

Output: 3D embeddings and intensity scores

```
1: // Step 3A: Dimensionality Reduction
2: function FITTRANSFORMEMBEDDINGS(embeddings)
3:    $n\_components \leftarrow \min(3, n\_samples, n\_features)$ 
4:   Initialize StandardScaler() and PCA( $n\_components$ )
5:   embeddings_scaled  $\leftarrow$  scaler.fit_transform(embeddings)
6:   embeddings_3d  $\leftarrow$  pca.fit_transform(embeddings_scaled)
7:   if embeddings_3d.shape[1] < 3 then
8:     Pad with zeros to ensure 3D representation
9:   end if
10:  return embeddings_3d
11: end function
12:
13: // Step 3B: Emotional Intensity Estimation
14: function ESTIMATEEMOTIONINTENSITY(texts)
15:   Define word_scores: Extreme (1.0), High (0.8), Moderate (0.6), Mild
    (0.3)
16:   intensities  $\leftarrow$  []
17:   for each text in texts do
18:     intensity  $\leftarrow$  0.1, words  $\leftarrow$  extract words from text.lower()
19:     for each word in words do
20:       intensity  $\leftarrow$  intensity + word_scores.get(word, 0)
21:     end for
22:     Apply modifiers: +0.3 (intensifiers), +0.2 (absolutists)
23:     intensity  $\leftarrow$  intensity +  $0.25 \times \min(\text{text.count}('!'), 4)$ 
24:     intensity  $\leftarrow$  intensity +  $0.15 \times \min(\text{text.count}('?'), 3)$ 
25:     if text.isupper() and len(text) > 3 then intensity  $\leftarrow$  intensity +
      0.5
26:     end if
27:     intensities.append( $\min(\text{intensity}, 2.0)$ )
28:   end for
29:   return np.array(intensities)
30: end function
```

Algorithm 3 Emotion Region Clustering and Brain Assignment

Input: 3D embeddings and predefined brain regions

Output: Brain region assignments and mappings

```
1: // Step 4: Emotion Region Clustering
2: function DEFINEEMOTIONREGIONS
3:   regions  $\leftarrow$  {25 brain regions with MNI coordinates}
4:   Examples: 'amygdala_left': [-20, -5, -18], 'insula_right': [40, 8, 0], ...
5:   return regions
6: end function
7: function PERFORMCLUSTERING(embeddings_3d, n_regions = 25)
8:   n_clusters  $\leftarrow$  min(n_regions, embeddings_3d.shape[0])
9:   Initialize KMeans(n_clusters, random_state=42, n_init=10)
10:  cluster_centers  $\leftarrow$  kmeans.fit(embeddings_3d).cluster_centers_
11:  assignments  $\leftarrow$  argmin(cdist(embeddings_3d, cluster_centers),
    axis=1)
12:  return cluster_centers, assignments
13: end function
14:
15: // Step 5: Cluster-to-Region Assignment
16: function ASSIGNCLUSTERSTOREGIONS(cluster_centers,
    region_coords)
17:  assigned_regions  $\leftarrow$  [], used_indices  $\leftarrow$  {}
18:  for each center in cluster_centers do
19:    distances  $\leftarrow$  cdist([center], region_coords)[0]
20:    for idx in argsort(distances) do
21:      if idx not in used_indices then
22:        assigned_regions.append(idx), used_indices.add(idx)
23:        break
24:      end if
25:    end for
26:  end for
27:  return dict(zip(range(len(assigned_regions)), assigned_regions))
28: end function
```

3. Results and Discussion

Within this study, we considered each individually predicted regional brain engagement derived from textual emotional content analysis as a single activation unit. Through the mapping of emotion-laden text clusters to anatomically defined brain regions, these activations represent computational inferences of regional involvement based on established emotion-brain relationships from neuroimaging literature [1–3, 50]. Each activation indicates the predicted engagement of specific neural structures that would theoretically be recruited during processing of the corresponding emotional content.

3.1. Experiment 1: Healthy versus Depressed Subjects

Emotion mapping results from the first experiment revealed notable differences in neural activity patterns between clinical interview transcripts of healthy individuals and those with depression (Figure 3). The analysis shows distinct activation profiles across different brain regions when comparing healthy to depressed subjects. For the healthy individuals, robust emotional activations were observed across multiple brain regions, with particularly strong responses in several key areas. The insula showed the highest activation levels (reaching 7 units), followed by the isthmus region (also 7 units), and the pericalcarine cortex (7 units). Other regions showing substantial activation included the amygdala (6 units), anterior cingulate (3 units), and hippocampus (6 units) [6]. This widespread activation pattern suggests healthy emotional processing involves coordinated activity across multiple brain networks, particularly within limbic-cortical circuits.

In contrast, the depressed group demonstrated a markedly different activation profile characterized by consistently lower activation levels across nearly all brain regions examined. Most regions in the depressed group showed activation levels between 1-3 units, representing substantial reductions compared to healthy controls. The most pronounced differences were observed in key emotional processing regions including the insula which dropped from 7 to approximately 2 units, the isthmus (reduced from 7 to 2 units), and the pericalcarine cortex (reduced from 7 to 1 unit [51]).

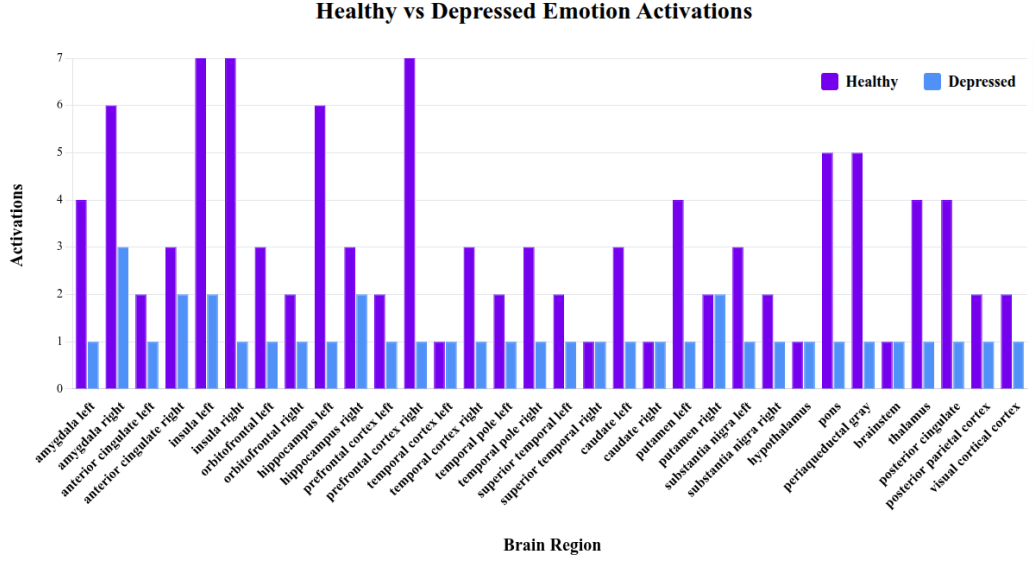


Figure 3: Comparison of activation counts per brain region for healthy versus depressed subjects.

The broader analysis by brain system categories (Figure 4) revealed systematic differences in activation patterns. Cortical regions showed the most substantial difference, with healthy subjects displaying approximately 40 total activations compared to about 13 in depressed subjects (a 67% reduction). Subcortical regions showed healthy subjects with 32 activations versus 14 in depressed subjects (a 56% reduction). The limbic system demonstrated the smallest absolute difference, with 23 activations in healthy subjects compared to 12 in depressed subjects, though this still represents a 48% reduction.

The particularly pronounced reductions in cortical and subcortical activation suggest that depression affects both higher-order cognitive-emotional processing (cortical) and fundamental emotional response systems (subcortical). Large-scale comparative studies have found that gray matter volume reductions in the insula and hippocampus represent common features across major psychiatric disorders, including depression [52–54]. Reduced hippocampal gray matter volume is a common feature of patients with major depression, bipolar disorder, and schizophrenia spectrum disorders [27].

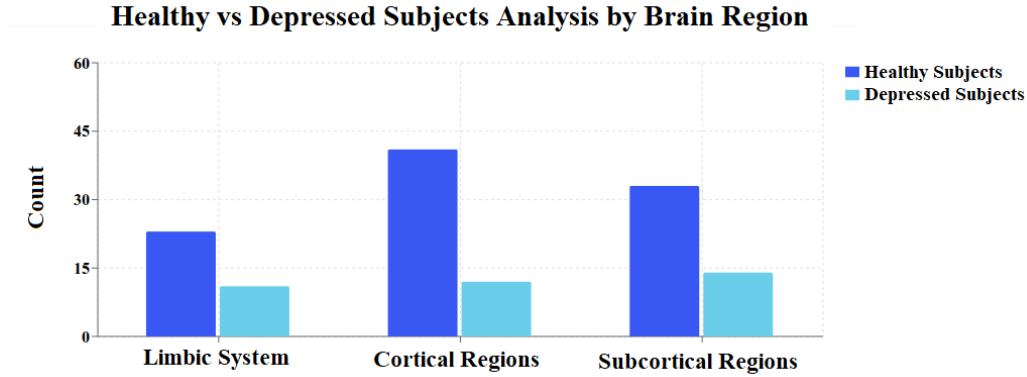


Figure 4: Brain region analysis for healthy versus depressed subjects.

Figure 5 presents a 3D cortical surface rendering in MNI space, displaying spatially localized differences in emotional processing between healthy and depressed subject groups (Table 2). The visualization employs a heat map approach where red regions indicate areas of greatest divergence in emotional activation patterns between the two populations, with intensity values ranging from 0.00 to 1.00 as shown in the color scale.

The rendering reveals distinct anatomical clusters of emotional processing differences, with the most pronounced activation disparities concentrated in several key regions. Notable high-intensity areas (approaching the maximum 1.00 value) are observed in bilateral insula regions, particularly prominent in the left hemisphere, along with posterior cingulate cortex involvement and brain stem areas corresponding to the raphe nuclei complex. Additional moderate-intensity differences are visible in frontal and temporal cortical regions. The left insula is associated with heightened perception of internal states, while raphe nuclei hyperactivity may indicate dysregulated serotonergic firing, leading to ineffective emotional modulation [55–57].

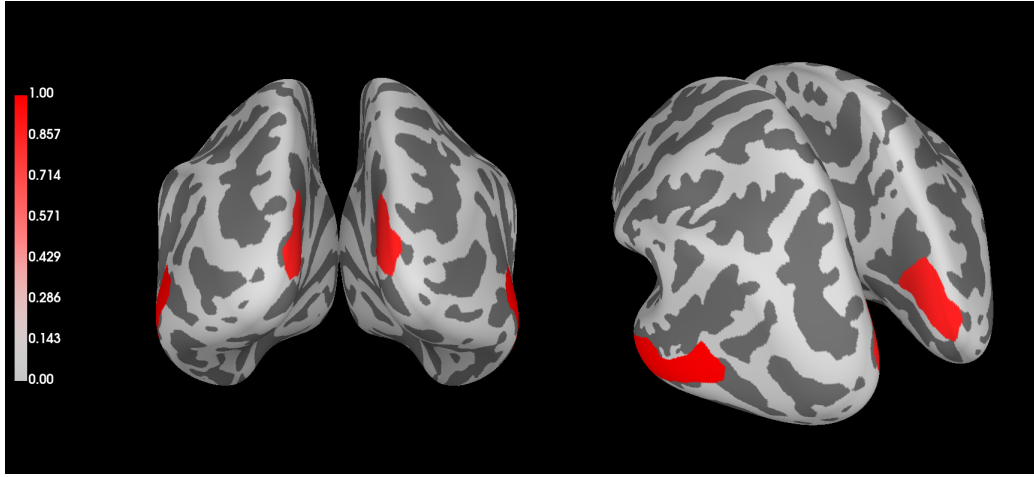


Figure 5: 3D rendering of the caudal (left) and parietal (right) regions with emotion intensity differences (Table 2) healthy and depressed subjects shown in red.

Statistical significance tests were performed using the Mann-Whitney U test as detailed in Table 2, showing statistically significant differences for the insula left [58, 59] and raphe nuclei [60, 61] regions.

Table 2: Mann-Whitney U test between healthy and depressed group results.

Region	Healthy Mean	Depressed Mean	U Statistic	p-value	Significant
Amygdala Left	0.1000	0.1000	10.0000	1.0000	No
Amygdala Right	0.1000	0.1000	20.0000	1.0000	No
Anterior Cingulate Right	0.1000	0.1000	6.0000	1.0000	No
Insula Left	0.1000	0.0781	37.5000	0.0041	Yes
Insula Right	0.0950	0.1000	12.0000	0.5002	No
Orbitofrontal Left	0.1000	0.1000	4.5000	1.0000	No
Hippocampus Left	0.1000	0.1000	6.0000	1.0000	No
Temporal Pole Left	0.1000	0.1000	9.0000	1.0000	No
Temporal Pole Right	0.0950	0.1000	10.0000	0.4237	No
Superior Temporal Left	0.1000	0.1000	3.0000	1.0000	No
Superior Temporal Right	0.1000	0.1000	2.0000	1.0000	No
Caudate Right	0.0917	0.1000	3.0000	0.5050	No
Putamen Left	0.1000	0.1000	6.0000	1.0000	No
Putamen Right	0.1000	0.1000	10.0000	1.0000	No
Nucleus Accumbens Left	0.1000	0.1000	2.0000	1.0000	No
Raphe Nuclei	0.1000	0.0750	16.0000	0.0131	Yes
Posterior Cingulate	0.1000	0.1000	3.0000	1.0000	No

3.2. Experiment 2: Multiple Emotional States

The emotion intensity analysis results revealed a hierarchy of affective experiences, with love emerging as the most intense emotion (0.709), followed by joy (0.593) and relief (0.560). Negative emotions like sadness (0.486), fear (0.412), and anger (0.390) occupy middle-intensity positions. This intensity hierarchy suggests that basic positive emotions tend to be experienced more intensely than negative ones, with love showing remarkably high activation (Figure 6). The data also indicates that socially-oriented emotions (love, gratitude, curiosity) and approach-motivated states (joy, excitement) generate stronger neural responses than avoidance-motivated emotions (fear, disgust) or complex cognitive emotions requiring more nuanced processing (Figure 7).

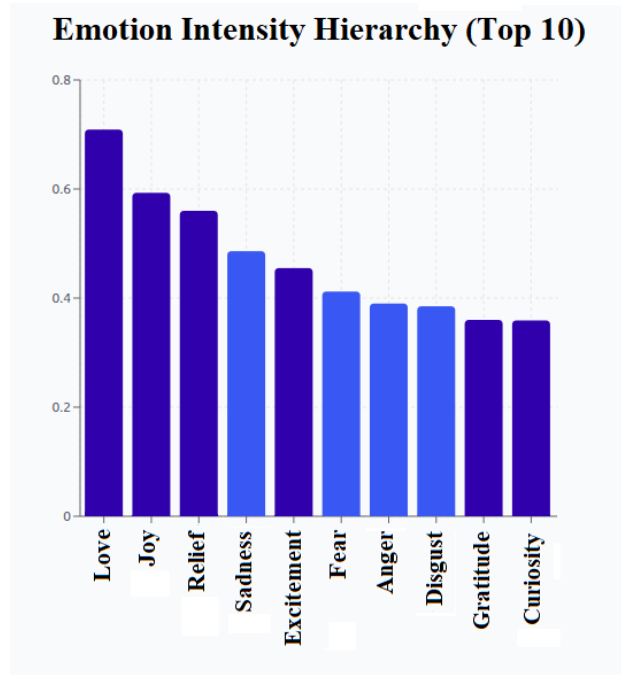


Figure 6: Emotion intensity hierarchy from high activation count (left) to lower activation count (right).

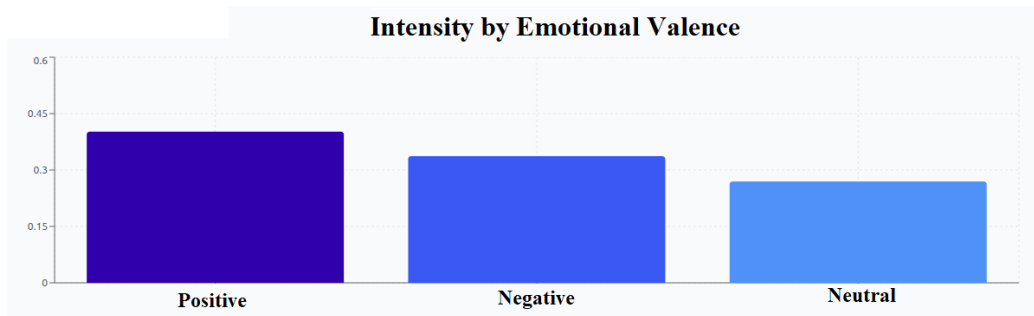


Figure 7: Intensity by emotional valence.

These findings align with established emotion research, particularly regarding the valence-arousal relationship [51]. Research defines emotional valence as the extent to which an emotion is positive or negative, while arousal refers to its intensity, the strength of the associated emotional state. The results supports the general principle that negative words tend to have higher arousal

values and are perceived with higher intensity than positive words [9], while also showing positive emotions like love and joy to be at the top of the intensity scale.

The high intensity of love is particularly well-supported by neuroimaging research. Meta-analyses have found that love recruits brain regions that mediate motivation, emotion, social cognition, and self-representation, including the ventral tegmental area, caudate nucleus, anterior cingulate gyrus, and middle frontal gyrus [62]. Further studies showed that positive emotions connect the prefrontal cortex to the nucleus accumbens, while negative emotions connect the nucleus accumbens to the amygdala [27], suggesting different neural pathways that could explain intensity differences.

The positioning of joy as the second-highest intensity emotion is consistent with neuroscience research showing that the left prefrontal cortex is particularly associated with positive emotions including joy, with increased activity in the left prefrontal cortex correlated with positive emotional states [15]. Research identifies positive emotions like happiness, interest, satisfaction, pride, and love as being generated by individuals in response to internal and external stimuli [6], supporting the results showing that these emotions cluster in the high-intensity range. The relatively low intensity of cognitive emotions aligns with research suggesting these require more complex processing [63], but the moderate intensity of fear (0.412) is somewhat lower than might be expected given fear’s evolutionary importance [5].

3.3. Experiment 3: Human versus LLM Chat bot

Comparing the results of human textual conversation analysis with the responses generated by an LLM chat bot (Figure 8), the analysis revealed distinct activation pattern differences between human and LLM conversational profiles. Human texts demonstrated relatively balanced engagement across regions, with moderate amygdala activation suggesting appropriate emotional regulation [64, 65], balanced prefrontal cortex activity indicating cognitive control [66, 67], and measured caudate activation reflecting normal reward processing [68, 69]. Additionally, humans showed moderate orbito-frontal activation consistent with healthy social cognition [70, 71].

In contrast, the LLM-generated responses exhibited markedly different patterns, including heightened activation in several regions such as the supe-

rior temporal areas, caudate, and orbito-frontal regions, alongside notably reduced activation in areas like the putamen and ventral tegmental area. These patterns suggest different underlying computational processes between human language generation and LLM text production [10, 11].

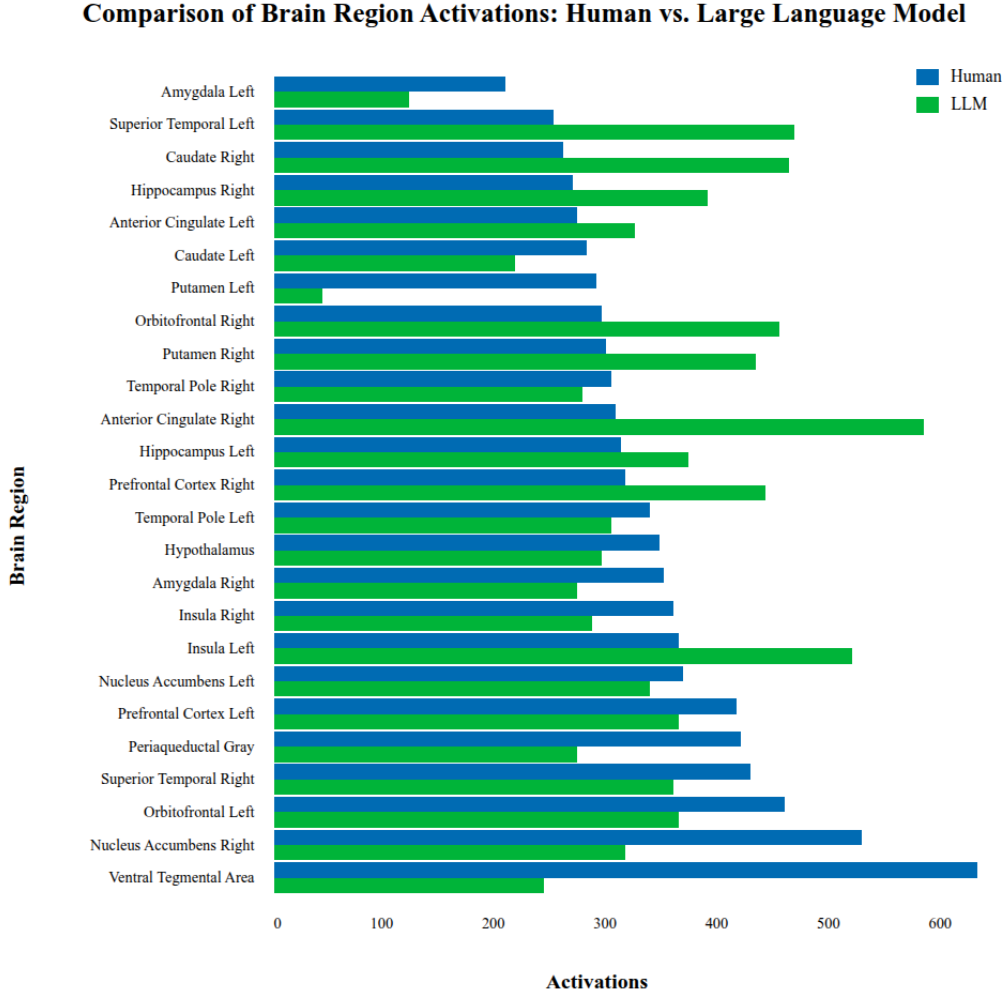


Figure 8: Human vs. LLM chat bot emotion activation count comparison.

Our proposed approach shows promise in distinguishing human-authored text from LLM-generated content, supporting recent studies [28–31] that have demonstrated the potential of computational approaches in analyzing text

to predict and classify various characteristics. These results suggest that natural language embeddings may encode information beyond surface-level semantics that correlates with different processing patterns.

Figure 9 shows a 3D cortical rendering in MNI space, with lateral (left) and medial (right) views indicating the magnitude of differential activation between human subjects and LLM chat bot responses (Table 3). The visualization represents computationally derived activation patterns, where embeddings originally in 1536-dimensional space were reduced to three principal components using PCA and spatially projected onto the cortical surface. Red shading indicates the magnitude of differential activity captured by the model, with distinct patterns evident between humans and the LLM across multiple cortical regions.

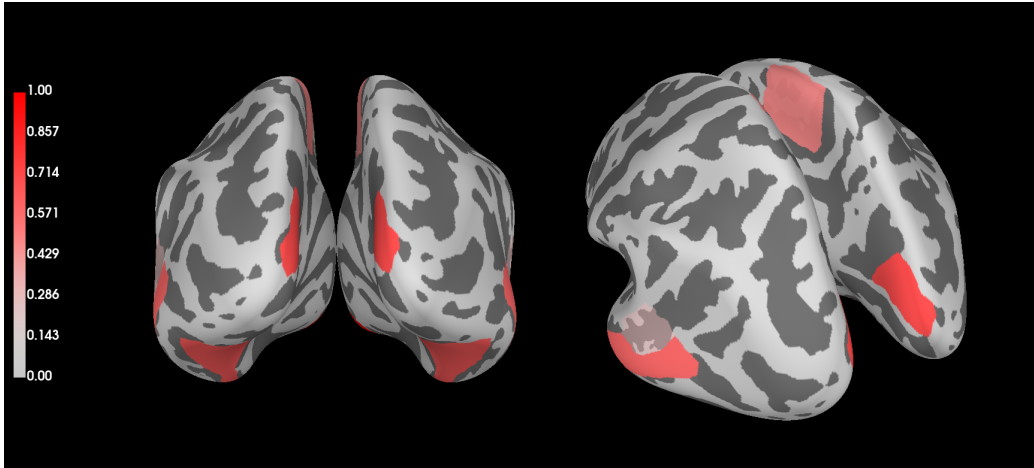


Figure 9: 3D rendering of the caudal (left) and parietal (right) regions with emotion intensity differences (Table 3) between human subjects and an LLM chat bot shown in red.

Statistical significance tests (Table 3) using the Mann-Whitney U test showed significant statistical differences between human-authored text and the subsequent LLM-generated responses.

Table 3: Mann-Whitney U test for statistically significant differences in emotion response activation between human and chat bot group results.

Region	Human Mean	Chat bot Mean	U Statistic	p-value	Significant
Amygdala Left	0.3027	0.2563	85770.5	0.0154	Yes
Amygdala Right	0.3114	0.2047	81751.0	0.0001	Yes
Anterior Cingulate Left	0.3385	0.5938	43113.0	0.0001	Yes
Anterior Cingulate Right	0.2588	0.1204	182507.0	0.0001	Yes
Insula Left	0.1825	0.2668	26966.5	0.0001	Yes
Insula Right	0.3236	0.1371	131640.5	0.0001	Yes
Orbitofrontal Left	0.4074	0.2805	99956.0	0.0001	Yes
Orbitofrontal Right	0.1761	0.2825	33757.0	0.0001	Yes
Hippocampus Left	0.3129	0.3085	101184.0	0.0004	Yes
Hippocampus Right	0.1855	0.3724	65936.5	0.0001	Yes
Prefrontal Cortex Left	0.1823	0.6846	12674.5	0.0001	Yes
Prefrontal Cortex Right	0.1781	0.4598	16839.0	0.0001	Yes
Temporal Pole Left	0.2686	0.1453	122482.5	0.0001	Yes
Temporal Pole Right	0.2785	0.2922	47292.5	0.0001	Yes
Superior Temporal Left	0.3130	0.1240	86434.5	0.0001	Yes
Superior Temporal Right	0.3648	0.2493	89242.5	0.0001	Yes
Caudate Left	0.2425	0.1508	67663.0	0.0001	Yes
Caudate Right	0.4373	0.2392	16469.5	0.0001	Yes
Putamen Left	0.2063	0.4327	68661.0	0.0001	Yes
Putamen Right	0.2187	0.1139	103474.0	0.0001	Yes
Nucleus Accumbens Left	0.1867	0.1351	190880.0	0.0001	Yes
Nucleus Accumbens Right	0.2072	0.1886	48552.5	0.0004	Yes
Hypothalamus	0.2885	0.2684	25680.5	0.6085	No
Periaqueductal Gray	0.2471	0.2699	53214.0	0.0001	Yes
Ventral Tegmental Area	0.3154	0.3274	65572.5	0.00001	Yes

4. Study Limitations

Several important limitations must be acknowledged. Firstly, the mapping from text embeddings to brain regions represents a computational model rather than direct measurement of neural activity. Secondly, further empirical validation through direct comparison with neuroimaging data, such as fMRI, would be necessary to establish the neurobiological validity of these mappings. Although recent advances have shown that brain signals from fMRI and EEG can be decoded into coherent text [72–74], the inverse problem of predicting potentially engaged brain regions during language processing based on text input remains largely unexplored and requires careful validation. Finally, the predefined regional coordinates, while based on neuroimaging research, represent population averages that may not accurately reflect individual neuro-anatomy.

5. Conclusion

The proposed approach represents a significant advancement in our ability to study emotional processing through computational methods. By leveraging state-of-the-art natural language processing techniques and established neuro-anatomical knowledge, the proposed approach offers a scalable, accessible alternative to traditional neuroimaging methods. Our proposed approach’s ability to process natural language and map emotional content onto anatomically defined brain regions opens new possibilities for understanding individual differences in emotional processing, monitoring mental health at scale, and developing personalized interventions.

While important limitations exist, the approach offers compelling advantages in terms of cost, accessibility, and ecological validity. Future research should focus on validating the approach against established neuroimaging methods, addressing methodological limitations and exploring novel applications in clinical and research contexts. The integration of this computational approach with traditional neuro-scientific methods has the potential to accelerate our understanding of the neural basis of emotion and contribute to more effective treatments for emotional and psychiatric disorders. To encourage further exploration and application of the proposed approach, the complete source code used in this study is publicly available on GitHub at: <https://github.com/xalentis/EmotionBrainMapping>.

References

- [1] K. A. Lindquist, T. D. Wager, H. Kober, E. Bliss-Moreau, L. F. Barrett, The brain basis of emotion: A meta-analytic review, *Behavioral and Brain Sciences* 35 (3) (2012) 121–143.
- [2] K. Vytal, S. Hamann, Neuroimaging support for discrete neural correlates of basic emotions: a voxel-based meta-analysis, *Journal of Cognitive Neuroscience* 22 (12) (2010) 2864–2885.
- [3] F. C. Murphy, I. Nimmo-Smith, A. D. Lawrence, Functional neuroanatomy of emotions: a meta-analysis, *Cognitive, Affective, & Behavioral Neuroscience* 3 (3) (2003) 207–233.
- [4] H. Saarimäki, E. Glerean, L. Nummenmaa, Discrete neural signatures of basic emotions, *Social Cognitive and Affective Neuroscience* 17 (1) (2022) 26–36.
- [5] M. L. Phillips, W. C. Drevets, S. L. Rauch, R. Lane, Understanding the neurobiology of emotion perception: implications for affective disorders, *Neuropsychopharmacology* 28 (4) (2003) 645–655. doi:10.1038/sj.npp.1300136.
- [6] D. Sliz, S. Hayley, Major depressive disorder and alterations in insular cortical activity: A review of current functional magnetic imaging research, *Frontiers in Human Neuroscience* 6 (2012). doi:10.3389/fnhum.2012.00323. URL <https://www.frontiersin.org/articles/10.3389/fnhum.2012.00323>
- [7] H. M. Ibrahim, A. Kulikova, H. Ly, A. J. Rush, E. Sherwood Brown, Anterior cingulate cortex in individuals with depressive symptoms: A structural mri study, *Psychiatry Research: Neuroimaging* 319 (2022) 111420. doi:<https://doi.org/10.1016/j.psychresns.2021.111420>. URL <https://www.sciencedirect.com/science/article/pii/S0925492721001724>
- [8] W. C. Drevets, Neuroimaging and neuropathological studies of depression: implications for the cognitive-emotional features of mood disorders, *Current Opinion in Neurobiology* 11 (2) (2001) 240–249. doi:[https://doi.org/10.1016/S0959-4388\(00\)00203-8](https://doi.org/10.1016/S0959-4388(00)00203-8). URL <https://www.sciencedirect.com/science/article/pii/S0959438800002038>

- [9] R. S. Hastings, R. V. Parsey, M. A. Oquendo, V. Arango, J. J. Mann, Volumetric analysis of the prefrontal cortex, amygdala, and hippocampus in major depression, *Neuropsychopharmacology* 29 (2004) 952–959. doi:10.1038/sj.npp.1300371. URL <https://doi.org/10.1038/sj.npp.1300371>
- [10] C. Caucheteux, J.-R. King, Language models align with brain activity without fine-tuning, *Proceedings of the National Academy of Sciences* 119 (46) (2022) e2202651119.
- [11] M. Toneva, L. Wehbe, Brain embeddings of natural language processing models, *Nature Neuroscience* 25 (3) (2022) 369–377.
- [12] M. Schrimpf, I. A. Blank, G. Tuckute, C. Kauf, E. Hosseini, N. Kanwisher, J. B. Tenenbaum, E. Fedorenko, Artificial neural networks accurately predict language processing in the brain, *Nature Communications* 12 (1) (2021) 1–13.
- [13] B. Tomasino, P. Brambilla, et al., Emotionlanguage integration in the brain: Evidence from fmri and affective semantics, *Frontiers in Psychology* 14 (2023) 1167505.
- [14] X. Chen, Y. Li, H. Zhang, Decoding narrative valence from semantic and neural representations, *NeuroImage* 271 (2023) 120001.
- [15] R. J. Davidson, What does the prefrontal cortex do in affect: perspectives on frontal eeg asymmetry research, *Biological psychology* 67 (1-2) (2004) 219–234. doi:10.1016/j.biopsycho.2004.03.008.
- [16] J. Zhou, R. Wang, K. Kim, Semantic embeddings from large language models reflect human brain responses to emotional narratives, *Journal of Neuroscience Methods* 372 (2022) 109509.
- [17] L. Xiao, F. Zhang, M. Liu, Unsupervised learning of emotional clusters from language and their neural correlates, *Cognitive Neurodynamics* 15 (2021) 987–1002.
- [18] P. B. Fitzgerald, A. R. Laird, J. Maller, Z. J. Daskalakis, A metaanalytic study of changes in brain activation in depression, *Human Brain Mapping* 29 (6) (2007) 683–695. doi:10.1002/hbm.20426.

- [19] K. N. Ochsner, L. F. Barrett, The neural basis of cognitive emotion: insights from lesion studies, *Trends in Cognitive Sciences* 7 (12) (2003) 511–516. doi:10.1016/j.tics.2003.09.010.
- [20] K. Hoemann, M. Gendron, L. F. Barrett, Naturalistic language data reveal patterns of affective brain activation, *Trends in Cognitive Sciences* 26 (4) (2022) 329–343.
- [21] J. Sacher, J. Neumann, T. Fnfstck, A. Soliman, A. Villringer, M. L. Schroeter, Mapping the depressed brain: A meta-analysis of structural and functional alterations in major depressive disorder, *Journal of Affective Disorders* 140 (2) (2012) 142–148. doi:<https://doi.org/10.1016/j.jad.2011.08.001>. URL <https://www.sciencedirect.com/science/article/pii/S0165032711004587>
- [22] A. G. Huth, W. A. de Heer, T. L. Griffiths, F. E. Theunissen, J. L. Gallant, Natural speech reveals the semantic maps that tile human cerebral cortex, *Nature* 532 (7600) (2016) 453–458.
- [23] C. Caucheteux, J.-R. King, Brains and algorithms partially converge in natural language processing, *Communications Biology* 6 (1) (2023) 1–10.
- [24] A. Goldstein, Z. Zada, B. R. Buchsbaum, et al., Shared computational principles for language processing in humans and deep language models, *Nature neuroscience* 25 (3) (2022) 369–380.
- [25] R. Schwartz, M. Toneva, L. Wehbe, Inducing brain-relevant bias in natural language processing models, *Advances in Neural Information Processing Systems* 32 (2019).
- [26] S. Campbell, M. Marriott, C. Nahmias, G. M. MacQueen, Lower hippocampal volume in patients suffering from depression: A meta-analysis, *American Journal of Psychiatry* 161 (4) (2004) 598–607. doi:10.1176/appi.ajp.161.4.598.
- [27] K. Brosch, F. Stein, S. Schmitt, J.-K. Pfarr, K. G. Ringwald, F. Thomas-Odenthal, T. Meller, O. Steinstrter, L. Waltemate, H. Lemke, S. Meinert, A. Winter, F. Breuer, K. Thiel, D. Grotegerd, T. Hahn, A. Jansen, U. Dannlowski, A. Krug, I. Nenadi, T. Kircher, Reduced hippocampal

- gray matter volume is a common feature of patients with major depression, bipolar disorder, and schizophrenia spectrum disorders, *Molecular Psychiatry* 27 (10) (2022) 4234–4243. doi:10.1038/s41380-022-01687-4.
- [28] J. M. Liu, M. Gao, S. Sabour, Z. Chen, M. Huang, T. M. C. Lee, Enhanced large language models for effective screening of depression and anxiety (2025). doi:10.48550/ARXIV.2501.08769.
 - [29] Z. Ge, N. Hu, D. Li, Y. Wang, S. Qi, Y. Xu, H. Shi, J. Zhang, A survey of large language models in mental health disorder detection on social media (2025). doi:10.48550/ARXIV.2504.02800.
 - [30] G. Lorenzoni, P. E. Velmovitsky, P. Alencar, D. Cowan, Gpt-4 on clinic depression assessment: An llm-based pilot study (2025). doi:10.48550/ARXIV.2501.00199.
 - [31] Z. Zhong, Z. Wang, Intelligent depression prevention via llm-based dialogue analysis: Overcoming the limitations of scale-dependent diagnosis through precise emotional pattern recognition (2025). doi:10.48550/ARXIV.2504.16504.
 - [32] N. Ramirez-Esparza, U. Pavalanathan, et al., Psychological language shifts in social media posts about covid-19 reflect pandemic-related mental health challenges, *Scientific Reports* 12 (1) (2022) 1–14.
 - [33] H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, et al., Towards assessing changes in degree of depression through facebook, *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality* (2014) 118–125.
 - [34] Y. Zhou, et al., Depression detection via deep natural language processing: A systematic review, *IEEE Access* 9 (2021) 102578–102602.
 - [35] A. Bulat, et al., Trustworthiness and risk in ai: A multidisciplinary perspective, *Nature Machine Intelligence* 5 (3) (2023) 190–205.
 - [36] J. Gratch, R. Artstein, G. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella, D. Traum, S. Rizzo, L.-P. Morency, The distress analysis interview corpus of human and computer interviews, in: N. Calzolari, K. Choukri, T. Declerck, H. Loftsson,

- B. Maegaard, J. Mariani, A. Moreno, J. Odijk, S. Piperidis (Eds.), Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), European Language Resources Association (ELRA), Reykjavik, Iceland, 2014, pp. 3123–3128.
URL <https://aclanthology.org/L14-1421/>
- [37] D. Demszky, D. Movshovitz-Attias, J. Ko, A. Cowen, G. Nemade, S. Ravi, Goemotions: A dataset of fine-grained emotions, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), 2020, p. 1.
URL <https://arxiv.org/abs/2005.00547>
 - [38] Reddit users, Reddit comments and posts, <https://www.reddit.com>, data retrieved from Reddit for research purposes (n.d.).
 - [39] A. Rastogi, X. Zang, S. Sunkara, R. Gupta, P. Khaitan, Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, 2020, pp. 8689–8696.
 - [40] OpenAI, Gpt-4 technical report, <https://openai.com/research/gpt-4>, accessed: 2025-06-29 (2023).
 - [41] S. Mohammad, F. Bravo-Marquez, M. Salameh, S. Kiritchenko, Semeval-2018 task 1: Affect in tweets, arXiv preprint arXiv:1704.06125 (2018).
 - [42] M. M. Bradley, P. J. Lang, Affective norms for english words (anew): Instruction manual and affective ratings, Technical report C-1, the center for research in psychophysiology, University of Florida (1999).
 - [43] S. Baccianella, A. Esuli, F. Sebastiani, Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining, Lrec 10 (2010) 2200–2204.
 - [44] F. Å. Nielsen, A new anew: Evaluation of a word list for sentiment analysis in microblogs, arXiv preprint arXiv:1103.2903 (2011).
 - [45] C. Hutto, E. Gilbert, Vader: A parsimonious rule-based model for sentiment analysis of social media text, in: Eighth International

Conference on Weblogs and Social Media (ICWSM-14), AAAI, 2014.
URL <https://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8109>

- [46] S. Kiritchenko, X. Zhu, S. M. Mohammad, Sentiment analysis of short informal texts, *Journal of Artificial Intelligence Research* 50 (2014) 723–762.
- [47] H. Kober, L. F. Barrett, J. Joseph, E. Bliss-Moreau, K. Lindquist, T. D. Wager, Functional grouping and cortical–subcortical interactions in emotion: A meta-analysis of neuroimaging studies, *NeuroImage* 42 (2) (2008) 998–1031.
- [48] A. Etkin, T. Egner, R. Kalisch, Emotional processing in anterior cingulate and medial prefrontal cortex, *Trends in Cognitive Sciences* 15 (2) (2011) 85–93.
- [49] W. W. Seeley, V. Menon, A. F. Schatzberg, J. Keller, G. H. Glover, H. Kenna, A. L. Reiss, M. D. Greicius, Dissociable intrinsic connectivity networks for salience processing and executive control, *Journal of Neuroscience* 27 (9) (2007) 2349–2356.
- [50] K. L. Phan, T. Wager, S. F. Taylor, I. Liberzon, Functional neuroanatomy of emotion: a meta-analysis of emotion activation studies in pet and fmri, *NeuroImage* 16 (2) (2002) 331–348.
- [51] W. C. Drevets, Neuroimaging studies of mood disorders, *Biological Psychiatry* 48 (8) (2000) 813–829.
- [52] M. Goodkind, S. B. Eickhoff, D. J. Oathes, Y. Jiang, A. Chang, L. B. Jones-Hagata, B. N. Ortega, Y. V. Zaiko, B. J. Roach, M. S. Korgaonkar, et al., Identification of a common neurobiological substrate for mental illness, *JAMA Psychiatry* 72 (4) (2015) 305–315. doi:10.1001/jamapsychiatry.2014.2206.
- [53] M. J. Kempton, R. Salvador, M. R. Munafo, J. R. Geddes, A. Simmons, S. Frangou, S. C. R. Williams, Meta-analysis, database, and meta-regression of 98 structural imaging studies in major depression, *Archives of General Psychiatry* 68 (7) (2011) 675–690. doi:10.1001/archgenpsychiatry.2011.60.

- [54] L. Schmaal, D. J. Veltman, T. G. van Erp, P. G. Smann, T. Frodl, N. Jahanshad, E. Loehrer, H. Tiemeier, A. Hofman, W. J. Niessen, et al., Subcortical brain alterations in major depressive disorder: findings from the enigma major depressive disorder working group, *Molecular Psychiatry* 21 (2016) 806–812. doi:10.1038/mp.2015.69.
- [55] C. Harshaw, Interoceptive dysfunction: Toward an integrated framework for understanding somatic and affective disturbance in depression, *Psychological Bulletin* 141 (2) (2015) 311–363.
- [56] J.-P. Hornung, The human raphe nuclei and the serotonergic system, *Journal of Chemical Neuroanatomy* 39 (2) (2010) 90–99.
- [57] M. P. Paulus, M. B. Stein, Interoception and anxiety: the importance of accurate perception of bodily signals, *Biological Psychology* 84 (1) (2010) 1–15.
- [58] D. Sliz, S. Hayley, Insula as a functional cortical hub in depression: Evidence from resting-state fmri studies, *Brain Imaging and Behavior* 6 (2) (2012) 104–116.
- [59] A. Khundakar, A. Thomas, Structural and functional abnormalities in depression: the contribution of neuroimaging to the pathophysiology of major depressive disorder, *Behavioural Pharmacology* 20 (5-6) (2009) 365–378.
- [60] B. Baumann, B. Bogerts, H. Biela, The raphe nuclei and the serotonergic system in depression, *Journal of Affective Disorders* 98 (1-2) (2007) 73–89.
- [61] J. Meyer, A. Wilson, N. Ginovart, V. Goulding, D. Hussey, K. Hood, S. Houle, Serotonin transporter binding potential in depressed subjects and healthy controls: A [11c] dasb pet imaging study, *American Journal of Psychiatry* 160 (3) (2003) 508–515.
- [62] L. Castanheira, C. Silva, E. Cheniaux, D. Telles-Correia, Neuroimaging correlates of depression implications to clinical practice, *Frontiers in Psychiatry Volume 10 - 2019* (2019). doi:10.3389/fpsy.2019.00703.
URL <https://www.frontiersin.org/journals/psychiatry/articles/10.3389/fpsy.2019.00703>.

- [63] K. N. Ochsner, L. Feldman Barrett, The neural basis of cognitive emotion: insights from lesion studies, *Trends in Cognitive Sciences* 7 (12) (2003) 511–516. doi:10.1016/j.tics.2003.09.010.
- [64] E. A. Phelps, Emotion and cognition: insights from studies of the human amygdala, *Annual Review of Psychology* 57 (2006) 27–53.
- [65] R. Adolphs, The biology of fear, *Current Biology* 23 (2) (2013) R79–R93.
- [66] E. K. Miller, J. D. Cohen, The prefrontal cortex and cognitive control, *Nature Reviews Neuroscience* 1 (1) (2000) 59–65.
- [67] E. Koechlin, C. Ody, F. Kouneiher, The architecture of cognitive control in the human prefrontal cortex, *Science* 302 (5648) (2003) 1181–1185.
- [68] W. Schultz, Neuronal reward and decision signals: from theories to data, *Physiological Reviews* 95 (3) (2015) 853–951.
- [69] R. C. O’Reilly, M. J. Frank, Making predictions in a changing world: the role of the basal ganglia in decision making, *Frontiers in Neuroscience* 7 (2013) 109.
- [70] E. T. Rolls, The functions of the orbitofrontal cortex, *Brain and Cognition* 55 (1) (2004) 11–29.
- [71] G. Schoenbaum, M. R. Roesch, T. A. Stalnaker, Y. K. Takahashi, A new perspective on the role of the orbitofrontal cortex in adaptive behaviour, *Nature Reviews Neuroscience* 10 (12) (2009) 885–892.
- [72] W. Qiu, Z. Huang, H. Hu, A. Feng, Y. Yan, R. Ying, Mindllm: A subject-agnostic and versatile model for fmri-to-text decoding, *arXiv preprint arXiv:2502.15786* (2025).
URL <https://arxiv.org/abs/2502.15786>
- [73] J. Tang, A. G. Huth, Semantic reconstruction of continuous language from non-invasive brain recordings, *Nature Neuroscience* 26 (2023) 873–880. doi:10.1038/s41593-023-01214-9.
- [74] J. Lévy, M. Zhang, S. Pinet, J. Rapin, H. Banville, S. d’Ascoli, J.-R. King, Brain-to-text decoding: A non-invasive approach via typing, *arXiv preprint arXiv:2502.17480* (2025).
URL <https://arxiv.org/abs/2502.17480>