

Large Language Models for Emotion Evolution Prediction

Clement Leung[1(✉)] and Zhifei Xu[2]

[1] School of Science and Engineering and Guangdong Provincial Key Laboratory of Future Networks of Intelligence, Chinese University of Hong Kong, Shenzhen, China
clementleung@cuhk.edu.cn

[2] School of Science and Engineering, Chinese University of Hong Kong, Shenzhen, China
zhifeixu@link.cuhk.edu.cn

Abstract. In numerous tasks, especially those of critical safety importance, it is essential that the human participants maintain appropriate emotional states. Recognizing these emotional states accurately has become a key focus, with mainstream approaches typically leveraging Pre-trained Language Models (PLMs) to integrate emotional insights. Recent advancements in Large Language Models (LLMs), like ChatGPT, have shown impressive proficiency in a range of natural language processing applications. Yet, the exploration into ChatGPT's zero-shot capabilities in the realm of image-based emotion recognition and analysis has been notably under-examined. In our research, we study the classification and prediction of emotions, distinguishing between positive and negative states, and we critically assess the ability of ChatGPT4 to recognize emotions from images, putting a spotlight on its potential in this area. The experimental results show that ChatGPT4 can effectively predict the evolution of emotion, and the accuracy of the evolution of positive and negative emotions is more than expected. However, ChatGPT4's prediction of specific negative emotions has gaps.

Keywords: image emotion prediction · Large Language Model · ChatGPT4 · zero-shot

1 Introduction and Related Works

In the complex landscape of human interactions, the nuanced portrayal and perception of emotions are fundamental. They not only facilitate meaningful connections but also serve as crucial indicators of well-being and intent. Meanwhile, research on how to imbue human-computer dialogue systems with emotions has It has been going on for ten years [3,4,12]. Furthermore, given the current perception of ChatGPT [14] and Instruct-GPT [18] as potent and versatile tools across various domains, we aim to explore their potential in facilitating accurate emotion recognition. Emotional support is increasingly becoming essential in numerous conversational contexts, including social interactions, mental

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2024
O. Gervasi et al. (Eds.): ICCSA 2024 Workshops, LNCS 14815, pp. 3–19, 2024.
https://doi.org/10.1007/978-3-031-65154-0_1

health assistance, and customer service chats. Thus, our research study assesses ChatGPT4's [1] efficiency and capability in identifying emotions through facial expressions.

Emotion recognition and prediction have garnered significant attention for their potential to enhance safety, support mental health, and improve user experiences. Emotion recognition and prediction have been recognized to be a significant factor affecting human safety and have been widely studied [2,27]. There exist in general multiple ways for people to express their emotions or feelings naturally, such as voice, text, video, facial expressions, and physical language.

In the modern world, where technology is ever more entwined with our daily activities, the ability of artificial intelligence to perceive and understand human emotions is becoming indispensable [22]. The importance of emotionally aware AI is magnified across various applications, reflecting the increasing pressures society faces today-ranging from criticism from superiors and unfair treatment to the pain of romantic separations [13,23]. Such stressors can push individuals to the brink, leading to a loss of emotional control. This, in turn, might result in harmful behaviors towards oneself or others. The evidence of emotional distress manifesting in severe outcomes is abundant: individuals grappling with suicidal ideation amidst academic or professional pressures [23]; the distressing prevalence of school shootings in the United States; or catastrophic accidents caused by road rage. Furthermore, in certain professions, emotional stability is critically important-consider the roles of surgeons, pilots, or truck drivers. For instance, there was a disturbing case where a pilot, suffering from depression, attempted to disable the plane's engines mid-flight [13]. This underscores the vital necessity for advancements in emotion recognition and prediction technologies in our lives [7].

In Recent years, generating emotional responses using neural networks has become a research hotspot [26], credit to the rapid development of online social networks and deep learning. In addition, with the continuous emergence of large language model technology and its rapid iterative development, many Human-computer interaction robots and artificial intelligence have brought a new technological revolution in the field of dialogue represented by ChatGPT4. At the same time, they demonstrate powerful general-purpose language processing capabilities and bring unprecedented human semantic understanding and response generation capabilities. Since their emergence, the interaction experience with human users has been greatly improved. However the question of whether emotions are shown in conversations has not yet been explored, and we are interested in developing effective conversation techniques in ChatGPT4. At the same time, we hope to explore the research progress, advantage analysis, and shortcomings [11,24,29] of ChatGPT4's multi-modal task emotion recognition and prediction fields. Moreover, the use of ChatGPT-4 for emotion recognition and prediction is deemed to enhance the fairness of experimental procedures. This is attributed to ChatGPT-4's ability as a large language model to analyze data without the bias and subjectivity often inherent in human judgment. The paragraph underscores the importance of this approach not only for maintaining fairness but

also for focusing on safety and well-being. By accurately identifying emotional states, interventions can be better tailored and more effectively implemented, thereby safeguarding participants and ensuring ethical standards are met within experimental frameworks.

In addition, in the field of emotion recognition and prediction, the main methods are studied by Ekman [5] and is one of the most famous classification models. He identified six basic emotions: joy, sadness, fear, anger, surprise, and disgust, which are considered universally identifiable. Plutchik has since expanded Ekman's model to propose eight major emotions [20] structured in a wheel format, including joy, trust, fear, surprise, sadness, disgust, anger, and anticipation. The model they proposed is a classification model, one of the two main framework models. Categorical models, also known as discrete emotion models, assume that there are a certain number of primary or basic emotions that are universally recognized and experienced by humans, regardless of cultural or individual differences. The other is the continuum model, which conceptualizes emotions as a spectrum or continuum [9,10], allowing for Valence (the positive or negative nature of the emotion), Arousal (the level of excitement or calmness), and Dominance (the level of excitement or calmness in a situation), degree of control or influence felt.

Studying emotion recognition and prediction has traditionally focused on a singular mode of expression. However, humans communicate emotions through various channels, including voice, text, video, facial expressions, and physical gestures, making it challenging to accurately discern emotions from a single source. The practice of multimodal sentiment classification, which merges information from multiple sources such as audio, dynamic shape, and global appearance features [17], addresses this complexity. This approach leverages the combined use of multi-modal datasets, incorporating both text and images, along with advanced analytical techniques like convolutional neural networks (CNNs) [8,28] or transforms for enhanced emotion recognition and classification. Given that emotions typically evolve in real-life situations, it is also crucial to predict their progression.

2 Methodology

Because in our real world, people's emotions are usually continuous, transitioning from one emotional state to another [15,21]. Depending on the actual situation, it is often necessary to predict and identify the emotional state of relevant personnel, for example: arranging surgeries in hospitals, pilots flying airplanes, scheduling work in high-risk industries, etc. As mentioned earlier, people may behave emotionally if they are tired or have been treated unfairly. In safety-critical jobs, we want the person doing the work to be in a sound emotional state [13]. In other words, people who work in high-risk industries can endanger the safety of others if they have emotional problems.

In this article, we focus on analyzing the classification model, while adopting the Ekman model, for Ekman's categories using the six primary emotions of

happiness, sadness, fear, anger, surprise, and disgust, we also group these into positive and negative emotions. We group the states happiness and surprise as positive (+1) emotions and group sadness, fear, anger, and disgust as negative (−1) emotions. As a variation, for safety-critical jobs, we may wish to be extra safe and more strict concerning positive emotion, and we may take surprise out from the positive emotion category, and place it in the negative emotion category. Here, though, we place surprise in the positive emotion category.

The model $S(t)$ encapsulates the emotional dynamics of an individual over time, where t represents the temporal dimension and $S(t)$ reflects shifts in emotional states due to time's passage. Specifically, the model assigns a value of 1 to denote positive emotions and -1 for negative emotions. The complexity of human emotions stems from an array of external stimuli that individuals have little to no control over. These stimuli can originate from numerous aspects of life such as financial status, personal relationships, health conditions, professional environments, the stock market, and family dynamics. It is the interplay of these factors that can prompt a shift from a state of positive to negative emotions, or the reverse.

Initially, we set $S(0) = 1$. Next, we model the moments of transition (switching from $+1$ to -1 or vice versa) using a Poisson Process. Consequently, $S(t)$ will equal 1 when the count of transitions within the period $(0, t)$ is an even number, and $S(t)$ will be -1 whenever this count is odd. Thus,

$$P[S(t) = 1|S(0) = 1] = p_0 + p_2 + p_4 + ... + ..., \tag{1}$$

where p_k is the number of Poisson points in $(0, t)$ with parameter λ. That is,

$$\begin{aligned} P[S(t) = 1|S(0) = 1] &= e^{-\lambda t}[1 + \frac{(\lambda t)^2}{2!} + \frac{(\lambda t)^4}{4!}... + ...] \\ &= e^{-\lambda t} \cosh \lambda t \end{aligned} \tag{2}$$

Now, $S(t) = -1$ if the number of points in the time interval $(0, t)$ is odd; that is,

$$\begin{aligned} P[S(t) = -1|S(0) = 1]] &= e^{-\lambda t}[1 + \frac{(\lambda t)^3}{3!} + \frac{(\lambda t)^5}{5!}... + ...] \\ &= e^{-\lambda t} \sinh \lambda t \end{aligned} \tag{3}$$

Equation 2 represents the probability that the emotion is still positive at time t given that it was positive at time 0. Equation 3 gives the probability that the emotion is negative at time t given that it was positive at time 0. The parameter λ in both expressions represents a rate at which emotions change or decay over time. A larger value of λ would mean emotions change more rapidly, while a smaller value would mean they change more slowly. This is where we mathematically analyze possible emotional changes and predict them. Also, to verify the idea, we use ChatGPT. As for the experimental evaluation part, the Receiver Operating Characteristic (ROC) method was adopted to analyze the experimental results, and a specific explanation was placed in the experimental part.

Here we briefly explain the above equation, first, we assume that lambda is 0.3, 0.6, 0.9. Meanwhile, The figure, labeled as Fig. 1, displays the function $e^{-\lambda t}\sinh(\lambda t)$ with λ values of 0.3, 0.6, and 0.9, Displaying the function $e^{-\lambda t}\sinh(\lambda t)$with λvalues of 0.3, 0.6, and 0.9, provides insightful observations about the temporal changes in probabilities, especially relevant to emotional states or comparable processes that may either diminish or progress over time. It is evident from the visualization that varying the decay constants (λ) significantly affects how long and intensely certain states persist across several days.

With a decay constant of $\lambda = 0.3$, the probability initially is high but diminishes gradually, indicating a persistent or slowly fading condition. For example, maintaining a negative emotional state translates to a higher likelihood of remaining in this state longer. Within the initial day, the probability stays well above 60%, denoting a strong endurance of the state. By day three, it hovers around 50%, showing a steady, yet noticeable decline. This gradual reduction might symbolize scenarios where the causes behind the emotional state are slow to be addressed or alleviated.

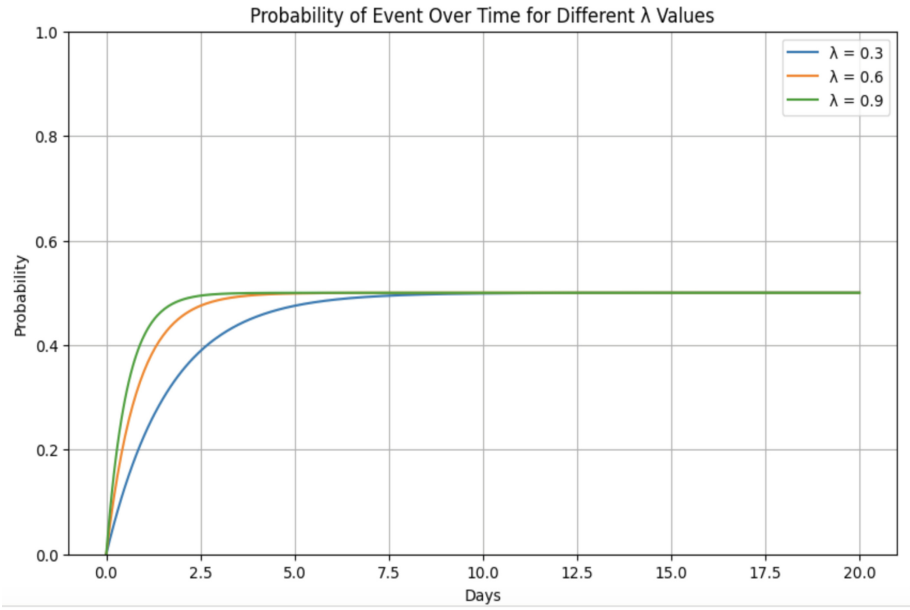

Fig. 1. Different λ of the Eq. 3

Increasing the decay constant to $\lambda = 0.6$ accelerates the probability's decline. This faster fall suggests a quicker fading of the emotional intensity or the likelihood of sustaining the same state. On the first day, the probability remains elevated but swiftly falls below 60%, nearing 50% by the close of the second day. This quicker reduction may be associated with effective interventions or environ-

mental changes, or possibly better coping strategies that shorten the duration of the negative condition.

At $\lambda = 0.9$, the probability decreases even more swiftly. The graph shows a sharp descent, indicative of scenarios where negative emotional states or similar conditions dissipate very quickly. The probability does not stay above 50% beyond two days, dropping near this mark by the end of the first day. This rapid decline could point to highly effective external support or events that inherently do not have prolonged effects.

By analyzing these curves, one can determine how various strategies or intrinsic elements affect the control or maintenance of specific states-be they emotional, physical, or of another nature. The differing λ values symbolically illustrate the varied speeds at which environments, individuals, or systems either normalize or transition from one state to another. This knowledge is essential in areas such as psychology, where predicting the duration of an individual's negative emotional state is key to developing timely and effective interventions. Insights into these temporal patterns are invaluable for customizing interventions or supports that are sensitive to timing and more closely correspond to the observed rates of change.

3 Experimentation

Emotion prediction in conversation is the convergence area of artificial intelligence and natural language processing. It involves analyzing data from text, images, and even audio to identify and predict the emotional state of participants during a conversation. This is useful for a variety of applications, such as improving customer service interactions, supporting mental health assessments, and enhancing human-computer interaction. In addition, conversational emotion recognition predictions can be analyzed and judged by chatbots to assess the current user's emotional state and responses to emotion drivers. From the above, ChatGPT4 is a form of conversational robot, and we hope to explore and analyze whether it can usefully predict changes in emotions.

3.1 Emotion Prediction with Different Situations

For the experimental part, we chose three Data sets from Kaggle which are Emotion Detection, Facial Expressions Training Data, and Natural Human Face Images for Emotion Recognition.

Dataset Emotion Detection. This dataset is the same as the FER-2013 [25] dataset. The collection features 35,685 grayscale images, each 48×48 pixels, organized into two sections: training and testing. Each section hosts a variety of images representing different emotional states. The images have been categorized by the creators into several emotions, namely anger, disgust, fear, happiness, neutrality, sadness, and surprise, providing a comprehensive basis for emotion detection tasks.

Table 1. Sample of Four Different Situations

Dataset	Question 1	Question 2	Question 3	Question 4
	what emotion of this person? If they are about to be praised by their boss or their parents respectively, what do you think their emotions become?	If they were to be criticized separately, what do you think their emotions would be?	If they were to each receive a $1,000 reward, what do you think their emotions would be?	If they were to break up, what do you think their emotions would be?
	what emotion of this person? If they are about to be praised by their boss or their parents respectively, what do you think their emotions become?	If they were to be criticized separately, what do you think their emotions would be?	If they were to each receive a $1,000 reward, what do you think their emotions would be?	If they were to break up, what do you think their emotions would be?
	what emotion of this person? If they are about to be praised by their boss or their parents respectively, what do you think their emotions become?	If they were to be criticized separately, what do you think their emotions would be?	If they were to each receive a $1,000 reward, what do you think their emotions would be?	If they were to break up, what do you think their emotions would be?

Facial Expression Training Data. The AffectNet [16] database, a substantial compilation of facial images annotated with expressions, serves as the foundation for this dataset. To adapt to typical memory constraints, image resolution is scaled down to 96×96 pixels. The dataset employs Principal Component Analysis, specifically focusing on the Singular Value Decomposition method, to enhance image processing efficiency. A threshold is applied to ensure the Principal Component's percentage remains below 90%, primarily excluding monochrome images. The dataset, derived from the high-quality AffectNet repository and refined using advanced Facial Expression Recognition technology, spans eight emotional categories: anger, contempt, disgust, fear, happiness, neutrality, sadness, and surprise.

Natural Human Face Images for Emotion Recognition. Unlike traditional datasets used in facial expression recognition like Facial Expression Recognition dataset (FER), Extended Cohn-Kanade dataset (CK+), and Karolinska Directed Emotional Faces dataset (KDEF), this unique dataset is curated from the internet, encompassing over 5,500 images manually labeled for eight emotional expressions: anger, contempt, disgust, fear, happiness, neutrality, sadness, and surprise. Each image, capturing real human expressions in 224×224 pixel grayscale format, is meticulously selected from various online sources, including Google, Unsplash, and Flickr, ensuring a wide array of natural facial expressions for enhanced learning and recognition tasks.

Task Definition of Emotion Prediction with Four Situations. According to the above description, we use three data sets and select 6 types anger, disgust, happiness, neutral, sadness, and surprise in the data set. In each data set, 10 images of 6 emotions are randomly selected and put into ChatGPT4 for judgment. As for the prompt words in Table 1, we want to preliminarily explore and predict the changes of emotion, so we choose four scenarios that are most likely to produce emotional changes in real life. At the same time, we artificially provide 4 situation simulations for each image, two positive situations, and two negative situations. (For details of specific questions see Table 1)

Predict the emotional changes of the image based on the simulated situation. Since ChatGPT4 will be released in 2023, the above experiments were all conducted using ChatGPT4. We use supervised learning and evaluate the performance of ChatGPT4 in a zero-shot prompt setting for the above tasks. After the judgment of ChatGPT4, if the result is the same as our cognitive result, it is recorded as 1, if the result is different, it is recorded as 0, in other words, the predicted results must be consistent with the logical results of most cognitive and emotional changes in real society and be consistent with common sense and recorded as positive or negative according to the emotion according to the description of ChatGPT4. Moreover, we construct a Receiver Operating Characteristic (ROC) [6] curve utilizing the outcomes we have documented. Within this curve, positive emotions such as happiness, neutrality, or surprise are assigned a value of 1, while negative emotions like anger, disgust, or sadness are designated with a value of 0. ChatGPT4's predictions for positive emotions are marked as 1 when they align with the actual outcomes, and as 0 when they do not. Similarly, for negative emotions, a matching prediction is indicated by a 0, and a mismatching one by a 1. The confidence level of these predictions is categorized on a scale from 1 to 3, where 1 indicates low confidence, 2 signifies moderate confidence, and 3 represents high confidence.

Table 2. Result of Four Different Situations

Emotion	Parameter	Positive Situation	Negative Situation
Anger	accuracy	68.30%	73.30%
	sensitivity	NaN	NaN
	specificity	68.30%	73.30%
Disgust	accuracy	78.30%	85.00%
	sensitivity	NaN	NaN
	specificity	78.30%	85.00%
Happiness	accuracy	91.70%	83.30%
	sensitivity	91.70%	83.30%
	specificity	NaN	NaN
Neutral	accuracy	86.70%	83.30%
	sensitivity	86.70%	83.30%
	specificity	NaN	NaN
Sad	accuracy	71.70%	80.00%
	sensitivity	NaN	NaN
	specificity	71.70%	80.00%
Surprise	accuracy	85.00%	90.00%
	sensitivity	85.00%	90.00%
	specificity	NaN	NaN
Negative	accuracy	72.80%	79.40%
	sensitivity	NaN	NaN
	specificity	72.80%	79.40%
Positive	accuracy	87.80%	85.60%
	sensitivity	87.80%	85.60%
	specificity	NaN	NaN

Preliminary Results. In the context of data presented in Table 2, the True Positive Rate (TPR), also referred to as Sensitivity, is a metric that quantifies the fraction of true positive instances accurately identified by the predictive model. Conversely, the False Positive Rate (FPR), also known as the complement of Specificity (1-Specificity), represents the proportion of negative cases that are mistakenly identified as positive by the model. The Observed Operating Points on the Receiver Operating Characteristic (ROC) curve signify the various thresholds applied within the classifier. Each of these points illustrates the equilibrium achieved between TPR and FPR at a given threshold setting. To elucidate, setting a higher threshold might lead to a reduction in FPR but

at the cost of diminishing TPR, whereas a lower threshold setting is likely to elevate both TPR and FPR. These critical points are instrumental in assessing the model's efficacy and in determining the optimal threshold for the task at hand, highlighting the inherent compromise between maximizing the detection of positive instances (achieving a higher TPR) and minimizing the occurrence of false positives (achieving a lower FPR).

Table 2 shows the prediction results of ChatGPT4 for the evolution of emotions after initially identifying negative and positive emotions and describing them through two positive situations and two negative situations respectively. For images initially identified as negative emotions, we found that their ChatGPT4 prediction accuracy in negative contexts was 79.4%. However, if the situation was positive, the predicted evolution of emotion was 72.8%. In contrast, for images initially identified as having positive emotions, their response accuracy was higher in positive contexts than in negative contexts. We think that it may be due to ChatGPT4, the possibility of negative emotions turning into positive emotions when encountering a positive environment is lower than the possibility of remaining negative in a negative environment. Positive emotions have the same result. This result shows that the prediction results of ChatGPT4 are consistent with the changes in our cognitive emotions since we want to preliminary explore and predict the changes in emotion, we choose four scenarios that are most likely to produce emotional changes in real life. For details in Table 2. The explanation is that the six categories of emotions were initially explored and analyzed separately, so when calculating the ROC, we also calculated the six categories of emotions separately to obtain the experimental results. According to the above description, positive emotions are recorded as 1 and negative emotions are recorded as 0. This means that when the six types of emotions are analyzed separately, they will lack the other half of the records. In other words, if the positive emotions are calculated as ROC When, it does not contain negative emotions, which is why there is a situation where specificity is *NaN*. Vice versa is also true.

Table 4 corresponds to the prediction results of emotional changes corresponding to different events that will occur under each different emotion. First, we preliminary observe that in the case of images depicting surprise or astonishment, GPT4 demonstrates a notable capability in recognizing these emotions as such. However, it encounters difficulty in discerning whether the surprise conveys a positive or negative sentiment, leading to a tendency to classify the emotion of surprise as predominantly neutral. Consequently, this is the reason why the outcomes for surprise closely mirror those associated with neutral expressions.

In order to avoid the harm caused by negative emotions to high-risk industries or high-risk groups, we mainly look at three types of emotions: anger, disgust, and sadness. We observe that in negative emotions, if the upcoming event is positive, then the accuracy of ChatGPT4 in predicting the emotion evolution

from high to low in zero-shot is disgust, sad, and anger; FPR is 78.3%, 71.7%, and 68.3% respectively. Anger is the strongest of negative emotions and the lowest in response to positive events. At the same time, because disgust is the most complex of negative emotions, including disgust, unhappiness, contempt, etc., it ranks the highest. Furthermore, the precision in identifying negative emotions falls short of expectations, suggesting that GPT4 could benefit from the inclusion of additional descriptive cues to enhance its decision-making process. Presently, in a zero-shot scenario, GPT4 is adept at recognizing the presence of negative emotions in individuals; however, it struggles with the accurate classification of specific emotions such as disgust, contempt, or anger. This is why negative predictions are less accurate than positive ones.

Analysis and Discussion. Throughout the training phase, it is common to encounter discrepancies between the emotions depicted in certain dataset images and our real-world perceptions. Due to the subjective nature of emotional interpretation, there is a possibility of encountering biases in recognizing the emotions conveyed by some images. In such instances, we rely on our judgment as the ultimate criterion and compare it to the interpretations provided by ChatGPT4.

Moreover, we have identified an additional complication: a misalignment between ChatGPT4's interpretations and the dataset's guidelines. A closer look at the specific examples of ChatGPT4's predictions highlights a fundamental issue-the disparity between the model's understanding and the dataset's standard. While the dataset might categorize an image as portraying anger based on its guidelines, ChatGPT4 might interpret the same expression as sadness or confusion. This discrepancy is not a matter of accuracy but rather an indication of differing standards used to classify negative emotions. Upon analysis, this divergence seems not solely a limitation of ChatGPT4 but could also stem from inadequate prompting. As the complexity of prompt instructions increases, expecting comprehensive coverage with minimal input becomes impractical. This realization opens up avenues for future improvements: if adhering strictly to the dataset's criteria is not mandatory, then refining the model based on broad prompt adjustments (like specifying the depicted emotions) might be viable. Yet, evaluating based on the dataset's labels could prove unsuitable, necessitating a more thorough manual review. On the contrary, if strict conformity to the dataset's guidelines is essential, relying on a multitude of prompt adjustments may fall short, making the supervised model fine-tuning a more effective strategy.

3.2 Emotion Prediction with Different Categories of Emotional Sentences

Dataset. First, we continue to use the same images as the previous task. They are still from emotion detection, facial expressions training data, and natural human faces. Each data set is still the same 10 images. But in the second task, we added a data set called MELD [19].

MELD. The Multimodal EmotionLines Dataset (MELD) builds upon and enriches the original EmotionLines dataset by incorporating additional modalities such as audio and visual elements alongside text. MELD features over 1,400 dialogue sequences and 13,000 spoken exchanges drawn from the "Friends" TV series, with various characters contributing to the conversations. Every piece of dialogue within MELD is categorized under one of seven possible emotions: Anger, Disgust, Sadness, Joy, Neutral, Surprise, and Fear. Additionally, MELD assigns a sentiment classification-positive, negative, or neutral-to each utterance, further enhancing its utility for emotion and sentiment analysis research.

Task Definition. The tasks in Part Two are partially similar to those in Part One. They all use the same images from the same dataset. However, each picture uses 6 categories of sentences full of different emotions 1. Anger, 2. Disgust, 3. Happiness, 4. Neutral, 5. Sad, 6. Surprise; think of these statements as what the character in the image is going to say. The input images and sentences are then analyzed using ChatGPT4 and the emotional evolution of any image is predicted and judged. (For details of specific questions see Table 3)

Preliminary Results. The abscissa of Table 5 represents the image of the dataset, and the ordinate represents the evolution of ChatGPT4's prediction of emotions after inputting 6 different emotional sentences. We can observe that the prediction accuracy of ChatGPT4 from high to low is: happiness, surprise, neutral, anger, sad, disgust. Among the three positive emotions, according to ChatGPT4 prediction, except for the happiness emotion that is directly converted into anger, which has the lowest accuracy, happiness is the highest for the others. At the same time, we observe that according to the description of ChatGPT4, when defining surprise and neutral, because they can be regarded as positive or negative, the results of the two are very similar. In the prediction of negative emotions, according to the above explanation of the FPR index, it shows that the disgust emotion is the least accurate to identify, and the emotion of the disgust category is the most difficult to judge among the six types of emotions. At the same time, still the same as the previous task, ChatGPT4 requires more prompts to achieve the accuracy of negative emotions. In the case of zero-shot, ChatGPT4 is not as good at predicting the evolution of emotions as in the case of positive emotions.

The vertical axis of the ROC curve denotes sensitivity, with its value directly proportional to the accuracy of diagnosis. Conversely, the horizontal axis signifies 1-specificity, where a lower value indicates a reduced rate of false positives. Generally, a point nearer to the upper left quadrant of the ROC space signals superior diagnostic performance, implying that a sensitivity approaching 1 correlates with enhanced predictive accuracy of the model. Before starting, we need to build on the explanation of partial definitions given earlier, this section

Table 3. Example of Six Different Categories Emotional Sentences

Dataset	Question 1	Question 2	Question 3	Question 4	Question 5	Question 6
	what emotion of these person? If the next thing they say is, "Well, why don't you tell her to stop being silly!" What do you think their emotions will become?	If the next sentence they say is, "Say it louder, I don't think the guy in the back heard you!" What do you think their emotions will become?	If the next sentence they say is, "Guess what, I got an audition!" What do you think their emotions will become?	If the next sentence they say is, "Great. He's doing great. Don't you worry about him?" What do you think their emotions will become?	If the next sentence they say is, "Yeah but we won't be able to like to get up in the middle of the night and have those long talks about our feelings and the future." What do you think their emotions will become?	If the next sentence they say is, "Look what I got! Look what I got! Can you believe they make these for little people?" What do you think their emotions will become?
	what emotion of these person? If the next thing they say is, "Well, why don't you tell her to stop being silly!" What do you think their emotions will become?	If the next sentence they say is, "Say it louder, I don't think the guy in the back heard you!" What do you think their emotions will become?	If the next sentence they say is, "Guess what, I got an audition!" What do you think their emotions will become?	If the next sentence they say is, "Great. He's doing great. Don't you worry about him?" What do you think their emotions will become?	If the next sentence they say is, "Yeah but we won't be able to like to get up in the middle of the night and have those long talks about our feelings and the future." What do you think their emotions will become?	If the next sentence they say is, "Look what I got! Look what I got! Can you believe they make these for little people?" What do you think their emotions will become?

focuses on the concepts of the Empirical ROC Area. The Empirical ROC Area, commonly known as the Area Under the Curve (AUC), quantifies the model's discriminative power directly from raw data by constructing an empirical ROC curve. This curve plots the True Positive Rate (TPR) against the False Pos-

Table 4. Result of Six Different Categories Emotional Sentences

Emotion	Anger sentence	disgust Sentence	Happiness sentence	Neutral Sentence	Sad sentence	Surprise sentence
Anger	70.00%	86.70%	86.70%	86.70%	86.70%	83.30%
Disgust	60.00%	70.00%	60.00%	56.70%	83.30%	56.70%
Happiness	70.00%	96.70%	1	96.70%	96.70%	96.70%
Neutral	76.70%	86.70%	96.70%	96.70%	90.00%	90.00%
Sad	63.30%	76.70%	76.70%	76.70%	86.70%	86.70%
Surprise	73.30%	86.70%	96.70%	96.70%	93.30%	96.70%

Table 5. Result of Dataset for Six Different Categories Emotional Sentences

Dataset	Parameter	Anger Sentence	Disgust Sentence	Happiness Sentence	Neutral Sentence	Sad Sentence	Surprise Sentence
Emotion Dection	accuracy	88.30%	53.30%	93.30%	90.00%	71.70%	91.70%
	sensitivity	83.30%	30.00%	96.70%	96.70%	70.00%	93.30%
	specificity	93.30%	76.70%	90.00%	83.30%	73.30%	90.00%
	Empiric ROC Area	0.989	0.837	0.997	0.994	0.92	0.993
Facial Expression	accuracy	81.70%	58.30%	93.30%	91.70%	78.30%	95.00%
	sensitivity	83.30%	46.70%	100	100	83.30%	96.70%
	specificity	80.00%	70.00%	86.70%	83.30%	73.30%	93.30%
	Empiric ROC Area	0.967	0.84	1	1	0.956	0.998
Neutral Human	accuracy	73.30%	58.30%	93.30%	93.30%	79.70%	85.00%
	sensitivity	76.70%	50.00%	100	100	79.30%	100
	specificity	70.00%	66.70%	66.70%	86.70%	80.00%	70.00%
	Empiric ROC Area	0.93	0.833	1	1	0.959	1

itive Rate (FPR) across a spectrum of decision thresholds. The AUC metric evaluates the model's efficacy in distinguishing between positive and negative classes' overall threshold values, with a larger AUC indicating superior model performance. An AUC value of 0.5 suggests no better than random class differentiation, while a value of 1.0 represents flawless discrimination. According to the table, we think that it may be due to the sensitivities of the three data sets being very close, except in the disgust statement. Because when the initial state has different emotions, it is difficult for ChatGPT4 to judge disgust. For example, in a positive situation, it will judge the disgusting statement as a joke or a prank, so the accuracy is not high. In terms of specificity, the prediction results of ChatGPT4 are beyond expectations, especially under the initial positive sentiment, the prediction accuracy is very high, almost close to being completely correct. Judging from the accuracy and ROC curve, ChatGPT4's performance in sentence prediction for different emotions exceeded expectations.

4 Conclusion

In this paper, we study the zero-shot capabilities of ChatGPT4 in sentiment reasoning and prediction based on multi-modal image text. The experimental results show that the prediction ability of ChatGPT4 is unexpectedly good and the accuracy is high. Since our goal is to analyze limitations, for example, ChatGPT4 can analyze that images have negative emotions, but it will incorrectly judge

disgust as depression, etc., resulting in inaccurate prediction results. However, with more prompt words and mental health analysis, ChatGPT4 has the ability and potential to improve inferences. Our findings indicate that tasks requiring subjective interpretation, like mental health assessments and understanding emotions in image-based conversations, pose significant challenges for ChatGPT. By enhancing the strategy for creating prompts and selecting relevant contextual examples, we posit that substantial further work is essential to boost ChatGPT's effectiveness and overcome its current shortcomings. This effort is crucial for its practical deployment in real-world scenarios about mental health and analogous fields.

5 Limitations and Future Work

In this paper, we evaluate sentiment primarily based on static inputs (images or single text inputs). In real-life scenarios, emotions are dynamic and can change rapidly based on ongoing interactions that current experimental setups may not fully capture. On the one hand, the lack of real-time feedback mechanisms to adjust model predictions based on user interactions limits the practical applicability of adaptive systems such as interactive chatbots or mental health monitoring tools. This article only focuses on the functions of ChatGPT4 in image-based emotion recognition. In the future, it may be extended to different large language models, such as Claude3, to discuss and compare the potential advantages and disadvantages of different large language models. On the other hand, an in-depth evaluation of performance in real-world scenarios is missing, raising questions about the robustness and generalization of the models beyond experimental settings. In the future, we can continue to explore and discuss how ChatGPT4 derives predictions or performs analytical observations by fine-tuning and adding prompt words, thereby improving the transparency and interpretability of the model's decision-making process at the same time. Judging based on emotional changes in perceived reality may bias the experiment. Since ChatGPT is a probabilistic model, even if the same result is given in the operation, it may give different results. Therefore, the next step might be to average the results over the dataset by running the same input multiple times, based on the disadvantage that a single experiment on the input is not sufficiently correlated.

References

1. OpenAI: GPT 4. https://openai.com/gpt-4
2. Ciraolo, D., Celesti, A., Fazio, M., Bonanno, M., Villari, M., Calabrò, R.S.: Emotional artificial intelligence enabled facial expression recognition for telerehabilitation: a preliminary study. In: 2023 IEEE Symposium on Computers and Communications (ISCC), pp. 1–6. IEEE (2023)
3. Deng, J.J., Leung, C.H.C.: Towards learning a joint representation from transformer in multimodal emotion recognition. In: Mahmud, M., Kaiser, M.S., Vassanelli, S., Dai, Q., Zhong, N. (eds.) BI 2021. LNCS (LNAI), vol. 12960, pp. 179–188. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-86993-9_17
4. Deng, J.J., Leung, C.H.C., Li, Y.: Multimodal emotion recognition using transfer learning on audio and text data. In: Gervasi, O., et al. (eds.) ICCSA 2021. LNCS, vol. 12951, pp. 552–563. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-86970-0_39
5. Ekman, P.: Facial expressions of emotion: New findings, new questions (1992)
6. Fawcett, T.: An introduction to ROC analysis. Pattern Recogn. Lett. **27**(8), 861–874 (2006)
7. Geetha, A., Mala, T., Priyanka, D., Uma, E.: Multimodal emotion recognition with deep learning: advancements, challenges, and future directions. Inf. Fusion **105**, 102218 (2024)
8. Issa, D., Demirci, M.F., Yazici, A.: Speech emotion recognition with deep convolutional neural networks. Biomed. Signal Process. Control **59**, 101894 (2020)
9. Kosti, R., Alvarez, J.M., Recasens, A., Lapedriza, A.: Emotion recognition in context. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1667–1675 (2017)
10. Kosti, R., Alvarez, J.M., Recasens, A., Lapedriza, A.: Context based emotion recognition using emotic dataset. IEEE Trans. Pattern Anal. Mach. Intell. **42**(11), 2755–2766 (2019)
11. Le, H.D., Lee, G.S., Kim, S.H., Kim, S., Yang, H.J.: Multi-label multimodal emotion recognition with transformer-based fusion and emotion-level representation learning. IEEE Access **11**, 14742–14751 (2023)
12. Leung, C.H., Deng, J.J., Li, Y.: Enhanced human-machine interactive learning for multimodal emotion recognition in dialogue system. In: Proceedings of the 2022 5th International Conference on Algorithms, Computing and Artificial Intelligence, pp. 1–7 (2022)
13. Lewis, R., Rose, J.: 'i'm not okay,' off-duty alaska pilot allegedly said before trying to cut the engines (OCTOBER 25, 2023, 11:55 AM ET). https://www.npr.org/2023/10/24/1208244311/alaska-airlines-off-duty-pilot-switch-off-engines
14. Mann, B., et al.: Language models are few-shot learners. arXiv preprint arXiv:2005.14165 (2020)
15. Metallinou, A., Narayanan, S.: Annotation and processing of continuous emotional attributes: challenges and opportunities. In: 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), pp. 1–8. IEEE (2013)
16. Mollahosseini, A., Hasani, B., Mahoor, M.H.: Affectnet: a database for facial expression, valence, and arousal computing in the wild. IEEE Trans. Affect. Comput. **10**(1), 18–31 (2017)

17. Nicolle, J., Rapp, V., Bailly, K., Prevost, L., Chetouani, M.: Robust continuous prediction of human emotions using multiscale dynamic cues. In: Proceedings of the 14th ACM International Conference on Multimodal Interaction, pp. 501–508 (2012)
18. Ouyang, L., et al.: Training language models to follow instructions with human feedback. Adv. Neural. Inf. Process. Syst. **35**, 27730–27744 (2022)
19. Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., Mihalcea, R.: Meld: a multimodal multi-party dataset for emotion recognition in conversations. arXiv preprint arXiv:1810.02508 (2018)
20. Robert, P.: Emotion: Theory, Research, and Experience, Volume 1: Theories of Emotion (1980)
21. SAMHSA: Warning signs and risk factors for emotional distress. https://www.samhsa.gov/find-help/disaster-distress-helpline/warning-signs-risk-factors
22. Shank, D.B., Graves, C., Gott, A., Gamez, P., Rodriguez, S.: Feeling our way to machine minds: people's emotions when perceiving mind in artificial intelligence. Comput. Hum. Behav. **98**, 256–266 (2019)
23. Fat cat incident. https://sports.sohu.com/a/776021122_121856967
24. Yang, K., Ji, S., Zhang, T., Xie, Q., Ananiadou, S.: On the evaluations of chatgpt and emotion-enhanced prompting for mental health analysis. arXiv preprint arXiv:2304.03347 (2023)
25. Zahara, L., Musa, P., Wibowo, E.P., Karim, I., Musa, S.B.: The facial emotion recognition (FER-2013) dataset for prediction system of micro-expressions face using the convolutional neural network (CNN) algorithm based raspberry PI. In: 2020 Fifth International Conference on Informatics and Computing (ICIC), pp. 1–9. IEEE (2020)
26. Zhang, R., Wang, Z., Huang, Z., Li, L., Zheng, M.: Predicting emotion reactions for human-computer conversation: a variational approach. IEEE Trans. Hum.-Mach. Syst. **51**(4), 279–287 (2021)
27. Zhang, T., Schoene, A.M., Ji, S., Ananiadou, S.: Natural language processing applied to mental illness detection: a narrative review. NPJ Digit. Med. **5**(1), 46 (2022)
28. Zhang, W., He, X., Lu, W.: Exploring discriminative representations for image emotion recognition with CNNs. IEEE Trans. Multimedia **22**(2), 515–523 (2019)
29. Zhao, W., Zhao, Y., Lu, X., Wang, S., Tong, Y., Qin, B.: Is chatgpt equipped with emotional dialogue capabilities? arXiv preprint arXiv:2304.09582 (2023)