Characterising Emotion Shifts Using Markov Processes

1st Clement Leung
School of Science and Engineering
Chinese University of Hong Kong
Shenzhen, China
clementleung@cuhk.edu.cn

2nd Zhifei Xu
School of Science and Engineering
Chinese University of Hong Kong
Shenzhen, China
zhifeixu1@link.cuhk.edu.cn

***Abstract*—In many operational contexts, particularly those that are safety-critical, it is imperative that human participants maintain appropriate emotional conditions. Consequently, the accurate recognition of these states is a central challenge in modern research. While mainstream methods have utilized Pre-trained Language Models (PLMs) for emotional understanding, the emergence of Large Language Models (LLMs) like ChatGPT offers new possibilities. This study investigates the underexplored zero-shot capabilities of ChatGPT-4 for image-based emotion analysis. We focus on its performance in classifying emotional valence (positive vs. negative) and predicting its temporal evolution. Our findings demonstrate that ChatGPT-4 can effectively forecast changes in emotional states, surpassing expectations. Nonetheless, we note deficiencies in its ability to accurately discern specific negative emotions, highlighting a need for further refinement. The study further introduces a hierarchical stochastic model to formalize these emotional shifts, providing a theoretical bridge between empirical LLM outputs and psychological stability parameters.**

Keywords-image emotion prediction; large language model; ChatGPT4; zero-shot; markov chain; emotion stability parameter.

I. Introduction

Accurately interpreting human emotion is fundamental to communication, enabling connection while revealing underlying mental states and intentions. For this reason, research has increasingly focused on integrating emotional insight into AI, from early human-computer dialogue systems [1][2] to the advanced Large Language Models (LLMs) of today. The arrival of models like ChatGPT [3] and Instruct-GPT [4] has sparked immense interest in LLM-based emotion recognition, particularly for providing emotional support in personal, clinical, and customer service settings. This study evaluates how effectively the latest iteration, ChatGPT-4 [5], can infer emotions from facial expressions alone.

The need for reliable emotion recognition is not merely academic; it is critical for safety, mental health, and user experience [6, 7]. Social stressors such as occupational strain, perceived injustice, and relationship loss can precipitate significant harm [8, 9]. Tragic incidents, including suicidal ideation linked to work demands [8], school shootings, road rage, and even a depressed pilot's attempt to shut down engines mid-flight [9], underscore the urgent need for better technological aids. Advanced emotion recognition and prediction systems could offer critical support for safety and mental health interventions [10].

While neural networks have long enabled emotionally responsive generation [11], the nuanced linguistic competence of modern LLMs like ChatGPT-4 has transformed conversational AI. Yet, the extent to which these systems can track or express emotion, especially through non-textual data, remains underexplored. This research assesses the strengths and limitations of ChatGPT-4 in multimodal emotion recognition and prediction [12, 13, 14]. By leveraging its capabilities, we can also reduce the human-rater bias often present in psychological studies, thereby promoting fairness and ethically tailored interventions.

A. Related Work: From Static to Generative Approaches

Historically, emotion recognition has relied on static classification models, such as Convolutional Neural Networks (CNNs) trained on fixed datasets like FER-2013 or AffectNet [15]. These "discriminative" models are excellent at categorizing a single frame but often fail to capture the temporal fluidity of human emotion. They view emotion as a snapshot rather than a process.

In contrast, Generative AI and LLMs offer a "generative" approach. They can synthesize context, history, and multimodal cues (text + image) to infer not just the current state, but the likely future state. However, the stochastic nature of LLMs introduces variability. This necessitates a robust mathematical framework to model that variability. Our work bridges this gap by applying stochastic process theory—specifically Markovian dynamics—to the output of generative models, providing a rigorous structure to the fluid predictions of an LLM.

Our work is grounded in established theories of emotion. These include categorical models, such as Ekman's six universal emotions (joy, sadness, fear, anger, surprise, disgust) [16] and Plutchik's wheel of eight (joy, trust, fear, surprise, sadness, disgust, anger, anticipation) [17], which posit a fixed set of basic emotions. In contrast, dimensional models view emotions along continuous axes of valence (positive/negative), arousal (intensity), and dominance [18, 19].

B. Contribution and Relation to Prior Work

This manuscript represents a substantial extension of our preliminary study presented at the BRAININFO 2025 conference [1]. While our initial work established the baseline feasibility of using ChatGPT-4 for zero-shot emotion prediction under hypothetical situations, the current study significantly expands the theoretical framework, experimental scope, and comparative analysis. The specific contributions that distinguish this article from the conference version are as follows:

1) **Hierarchical Stochastic Modeling:** We upgrade the mathematical framework from a standard Markov chain to a hierarchical model. This includes the introduction of a binary valence layer based on a Poisson process (Section II-B), which mathematically links global emotional volatility to categorical transitions.
2) **Multimodal Dataset Expansion:** Whereas [1] relied exclusively on static facial expression datasets, this study incorporates the Multimodal EmotionLines Dataset (MELD). This allows us to evaluate the model's performance on complex scenarios involving dialogue and sentiment-tagged utterances.
3) **New Experimental Tasks:** We introduce a new prediction task involving emotion-conditioned sentences. Unlike the situational prompts used in [1], this task tests the model's ability to predict emotional evolution based on specific verbal cues (e.g., an angry utterance vs. a surprised utterance).
4) **Comparative Analysis:** We provide a comprehensive comparison between ChatGPT-4 and the Doubao (TikTok) Large Language Model, highlighting critical divergences in how these models interpret negative emotional states and zero-shot multimodal prompts.

Section II introduces the hierarchical stochastic model used to formalise emotion shifts. Section III describes the datasets, prompting protocol, and quantitative evaluation results. Section IV discusses limitations, ethical considerations, and future directions.

C. Problem Setting and Research Questions

We study *zero-shot* emotion inference where the model receives (i) a facial image and (ii) an optional textual continuation (a scenario description or an emotion-conditioned utterance), and must output both a current emotion label and a plausible next emotion label. This differs from standard facial-expression classification in two ways. First, the output is inherently *temporal* (a transition rather than a single label). Second, the "ground truth" for a hypothetical future emotion is not directly observable; therefore, our evaluation separates (a) *recognition correctness* (agreement with dataset labels for the current frame) from (b) *transition consistency* under controlled polarity cues (positive vs. negative situations) and under utterances drawn from MELD-style emotion categories.

Accordingly, we structure the study around three research questions:

- **RQ1 (Recognition):** When prompted with facial images only, how reliably can ChatGPT-4 infer the dataset emotion label, and how does performance differ between positive vs. negative categories?
- **RQ2 (Shift prediction):** Given an initial facial emotion, does the model predict transitions that are *consistent* with the polarity of the subsequent situation/utterance (e.g., reward-like vs. breakup-like contexts), and where does it fail?
- **RQ3 (Mechanism):** Can a compact stochastic process model (Poisson + Markov + persistence) explain the empirical pattern that valence is often correct while fine-grained negative categories are frequently confused?

These questions motivate our hierarchical model in Section II and the prompting/evaluation protocol in Section III.

II. Mathematical Model

This section formalizes the stochastic model that we use to describe the temporal evolution of emotions and to interpret the empirical behaviour of ChatGPT-4 and Doubao in Section III. The construction proceeds in three layers: (i) a binary valence layer based on a Poisson process, (ii) a categorical layer using an eight-state Markov chain, and (iii) a stability layer with emotion-specific persistence parameters.

A. Notation

Table I summarises the main notation used in this section.

B. Binary valence model (Poisson switching)

At the coarsest level, we distinguish positive from negative valence. Let

$$S(t) \in \{+1, -1\} \tag{1}$$

denote the valence state at continuous time t, with $S(0) = +1$ indicating an initially positive state.

Valence switches are driven by a homogeneous Poisson process $N(t)$ with rate $\lambda > 0$. Each arrival of the process flips the sign of $S(t)$. If the number of arrivals in $(0, t]$ is even, the valence remains positive; if it is odd, the valence is negative.

Let $p_k = \Pr\{N(t) = k\}$ be the Poisson probabilities with parameter λt. The probability that valence is still positive at time t, given that it started positive, is

$$\Pr\{S(t) = 1 \mid S(0) = 1\} = p_0 + p_2 + p_4 + \cdots = e^{-\lambda t}\cosh(\lambda t). \tag{2}$$

Similarly, the probability that the state has flipped to negative is

$$\Pr\{S(t) = -1 \mid S(0) = 1\} = e^{-\lambda t}\sinh(\lambda t). \tag{3}$$

The parameter λ therefore acts as a global emotional volatility parameter: small λ implies long-lasting valence (rare switches), whereas large λ produces rapid alternation between positive and negative states.

1) Discrete-step interpretation and an explicit stay/flip form: In many applications the model is queried at discrete steps (e.g., turns in a dialogue or time bins of a fixed duration Δt). Under Poisson-driven sign flips, the probability of *staying* in the same valence over one step is

$$\Pr\{S(t+\Delta t) = S(t)\} = e^{-\lambda \Delta t}\cosh(\lambda \Delta t) = \frac{1 + e^{-2\lambda \Delta t}}{2}, \tag{4}$$

and the probability of a *flip* is

$$\Pr\{S(t+\Delta t) \neq S(t)\} = e^{-\lambda \Delta t}\sinh(\lambda \Delta t) = \frac{1 - e^{-2\lambda \Delta t}}{2}. \tag{5}$$

TABLE I
MAIN NOTATION USED IN THE MODEL.

Symbol	Description
$S(t)$	Valence state at continuous time t (+1 = positive, -1 = negative)
$N(t)$	Poisson process counting valence switches up to time t
λ	Global valence switching rate (Poisson intensity)
E_t	Categorical emotion at discrete step t
E	Emotion set $\{$Joy, Trust, Surprise, Anticipation, Sadness, Disgust, Anger, Fear$\}$
E_+	Positive emotions $\{$Joy, Trust, Surprise, Anticipation$\}$
E_-	Negative emotions $\{$Sadness, Disgust, Anger, Fear$\}$
$p_i(t)$	Probability $\Pr\{E_t = E_i\}$ of emotion E_i at step t
$\mathbf{p}(t)$	Column vector $[p_1(t), \ldots, p_8(t)]^\top$
$\tilde{p}_i(t)$	Stability-adjusted probability of emotion E_i at step t
λ_i	Stability parameter for emotion E_i (smaller = more persistent)
P_{ij}	One-step transition probability from E_i to E_j
P	8×8 row-stochastic state transition matrix

These closed forms clarify how λ controls volatility: for small $\lambda\Delta t$, flips are rare; as $\lambda\Delta t$ grows, the process approaches a near-random alternation with stay probability $\approx 1/2$.

Moreover, if an empirical estimate $\widehat{p}_{\text{stay}}$ of the valence stay probability over Δt is available, one may invert (4) to obtain

$$\widehat{\lambda} = -\frac{1}{2\Delta t}\ln\big(2\widehat{p}_{\text{stay}} - 1\big), \quad \text{valid when } \widehat{p}_{\text{stay}} > \tfrac{1}{2}. \tag{6}$$

This provides a principled link between observed stability (from repeated LLM trajectories) and the volatility parameter.

C. Categorical extension: eight-emotion Markov chain

To represent which emotion is being expressed, we refine the valence layer into eight categorical states,

$$E = \{\text{Joy}, \text{Trust}, \text{Surprise}, \text{Anticipation}, \tag{7}$$

$$\text{Sadness}, \text{Disgust}, \text{Anger}, \text{Fear}\}. \tag{8}$$

We partition these into positive and negative subsets,

$$E_+ = \{\text{Joy}, \text{Trust}, \text{Surprise}, \text{Anticipation}\} \tag{9}$$

$$E_- = \{\text{Sadness}, \text{Disgust}, \text{Anger}, \text{Fear}\} \tag{10}$$

and define a simple valence map $g : E \to \{+1, -1\}$ with $g(E_i) = +1$ for $E_i \in E_+$ and $g(E_i) = -1$ for $E_i \in E_-$.

Time is now indexed in discrete steps $t \in \{0, 1, 2, \ldots\}$ (e.g., conversational turns or fixed-size time bins). Let E_t denote the emotion at step t, and define

$$p_i(t) = \Pr\{E_t = E_i\}, \qquad \mathbf{p}(t) = [p_1(t), \ldots, p_8(t)]^\top, \tag{11}$$

with $\sum_{i=1}^{8} p_i(t) = 1$.

The categorical dynamics follow an eight-state Markov chain with transition matrix P:

$$P_{ij} = \Pr\{E_{t+1} = E_j \mid E_t = E_i\}, \qquad \sum_{j=1}^{8} P_{ij} = 1 \quad \forall i. \tag{12}$$

Using the column-vector convention, the one-step update is

$$\mathbf{p}(t) = P^\top \mathbf{p}(t-1). \tag{13}$$

1) Theoretical Implications: This hierarchical structure implies that emotional stability is not uniform. The Poisson layer dictates the "mood" (valence), while the Markov layer dictates the specific "affect" (emotion). This aligns with psychological appraisal theories where a general valence check often precedes specific emotional labeling. In our experiments with ChatGPT-4, we observe that the model often gets the valence correct (Poisson layer) even when it confuses the specific category (Markov layer), supporting the validity of this hierarchical separation.

D. Stability and persistence parameters

To keep the model simple and interpretable, we group emotions by polarity and assign

$$\lambda_i = \begin{cases} 0.2, & E_i \in E_+ \quad \text{(more persistent positive emotions)}, \\ 0.5, & E_i \in E_- \quad \text{(more volatile negative emotions)}. \end{cases} \tag{14}$$

Given a current distribution $\mathbf{p}(t) = [p_1(t), \ldots, p_8(t)]^\top$, the probability that emotion E_i stays the same at time t is modelled analogously to (2):

$$P_{\text{stay},i}(t) = p_i(t)\, e^{-\lambda_i t} \cosh(\lambda_i t). \tag{15}$$

The complementary probability mass $p_i(t) - P_{\text{stay},i}(t)$ corresponds to transitions out of E_i.

We then redistribute this transition mass according to the matrix P. Let P_{ji} be the probability of moving from E_j to E_i. The stability-adjusted probability of emotion E_i at time t is

$$\tilde{p}_i(t) = P_{\text{stay},i}(t) + \sum_{j \neq i} \big[p_j(t) - P_{\text{stay},j}(t)\big]\, P_{ji}. \tag{16}$$

E. Constructing P from empirical transitions

The Markov transition matrix P can be interpreted in two complementary ways. First, it can be treated as a *theoretical prior* encoding psychologically plausible shifts (e.g., Surprise → Joy under positive contexts). Second, it can be estimated from model-generated trajectories to summarise how a particular LLM tends to "move" between emotion labels.

Concretely, suppose we collect C_{ij} counts of predicted one-step transitions $E_t = E_i \to E_{t+1} = E_j$ across all prompts and samples. A maximum-likelihood estimate is obtained by row-normalising:

$$\widehat{P}_{ij} = \frac{C_{ij}}{\sum_{k=1}^{8} C_{ik}}. \tag{17}$$

To avoid zero-probability artifacts (common when some transitions are rarely observed), a simple additive smoothing scheme can be used:

$$\widehat{P}_{ij}^{(\alpha)} = \frac{C_{ij} + \alpha}{\sum_{k=1}^{8} (C_{ik} + \alpha)}, \tag{18}$$

where $\alpha > 0$ acts like a symmetric Dirichlet prior and guarantees a well-defined stochastic matrix. In Section III, we primarily use P to (i) generate reference trajectories via Algorithm 1 and (ii) interpret confusion patterns: large off-diagonal mass from a negative emotion into Neutral/Joy-like predictions is consistent with low specificity and reduced AUC for that category.

F. Numerical Simulation Algorithm

To visualize the prediction process, we formalize the simulation steps in Algorithm 1. This algorithm iteratively updates the emotion state vector based on the Markov transition matrix and stability adjustments defined above.

Algorithm 1: Emotion Evolution Simulation

```
Input: Initial state vector $\mathbf{p}(0)$, Transition Matrix $P$,
  Stability parameters $\lambda_i$, Time horizon $T$.
Output: Probability distributions $\tilde{\mathbf{p}}(t)$ for $t = 1 \dots T$.
for $t = 1$ to $T$ do
    // Step 1: Standard Markov Update
    $\mathbf{p}(t) \leftarrow P^{\top}\mathbf{p}(t-1)$;
    // Step 2: Calculate Persistence
    for $i = 1$ to $8$ do
        $P_{\text{stay},i}(t) \leftarrow p_i(t) e^{-\lambda_i t} \cosh(\lambda_i t)$;
    end
    // Step 3: Redistribute Mass
    for $i = 1$ to $8$ do
        $\tilde{p}_i(t) \leftarrow P_{\text{stay},i}(t) + \sum_{j \neq i} [p_j(t) - P_{\text{stay},j}(t)] P_{ji}$;
    end
    // Step 4: Normalize and Store
    $\tilde{\mathbf{p}}(t) \leftarrow \text{Norm}(\tilde{\mathbf{p}}(t))$;
end
return $\tilde{\mathbf{p}}(1 \dots T)$
```

This algorithmic approach ensures that for any given initial emotion detected by the LLM, we can project a probabilistic trajectory of how that emotion might decay or shift, providing a benchmark to compare against the LLM's own predictions.

G. Connection to ROC/AUC metrics and LLM experiments

The model above provides a conceptual bridge between emotional stability and the classification metrics observed in Section III. At the valence level, a larger global λ or larger negative-emotion λ_i produces more frequent sign flips and greater overlap between positive and negative trajectories. In classical detection theory, increased overlap translates into lower AUC: the ROC curve moves closer to the diagonal.

Empirically, we observe that positive emotions (e.g., happiness, surprise) achieve high accuracies and AUC values close to 1, indicating stable, well-separated positive trajectories. Negative emotions, especially disgust, exhibit lower accuracies and smaller AUC, suggesting that their score distributions overlap more with positive classes. This pattern is precisely what the model predicts when negative emotions have larger λ_i (more volatile, shorter dwell times).

III. Experimental Design and Results

Understanding and predicting emotion is a major frontier in conversational AI. By analyzing not just the words people use, but also visual and auditory cues, we can forecast how their feelings will shift throughout a dialogue.

A. Evaluation Metrics

To rigorously assess the model's performance, we utilize standard classification metrics derived from the confusion matrix. Let TP be True Positives, TN be True Negatives, FP be False Positives, and FN be False Negatives.

- **Accuracy:** The proportion of total predictions that are correct.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{19}$$

- **Sensitivity (Recall):** The ability of the model to correctly identify positive emotional states.

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{20}$$

- **Specificity:** The ability of the model to correctly identify negative emotional states.

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{21}$$

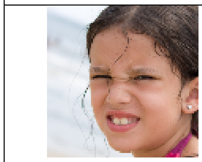
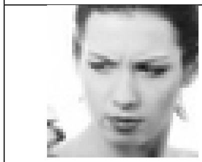
Additionally, we calculate the Area Under the Receiver Operating Characteristic Curve (AUC-ROC), which plots Sensitivity against $1 -$ Specificity. An AUC of 0.5 represents random guessing, while 1.0 represents perfect classification.

1) Uncertainty reporting: Point estimates can hide variability across samples and prompts. Where space permits, we recommend reporting uncertainty via nonparametric bootstrap confidence intervals. Specifically, we resample the evaluation set with replacement, recompute Accuracy and AUC for each resample, and report the 2.5/97.5 percentiles as a 95% interval. This is particularly important when comparing models (ChatGPT-4 vs. Doubao) where differences may be concentrated in a small subset of hard negative categories.

B. Emotion Recognition 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.

TABLE II
SAMPLE OF FOUR DIFFERENT SITUATIONS

Dataset	Question 1	Question 2	Question 3	Question 4
	What is the 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, what do you think their emotions would be?	If they were to 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 is the 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, what do you think their emotions would be?	If they were to 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 is the 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, what do you think their emotions would be?	If they were to 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?

1) Label harmonisation across datasets and the eight-state model: Different datasets use partially overlapping taxonomies. For consistent reporting, we focus on the shared labels {anger, disgust, happiness, neutral, sadness, surprise} for the six-way experiments. Our stochastic model uses an eight-state affect set inspired by Plutchik; the mapping is summarised in Table III. Neutral is treated as a separate category in evaluation (not one of the eight affect states), which is a common practical compromise when combining categorical theories with "no strong affect" dataset labels.

2) Datasets: **Emotion Dection** This dataset is the same as the FER-2013 [20] dataset. The collection features 35,685 grayscale images, each 48x48 pixels. The images have been categorized by the creators into several emotions, namely anger, disgust, fear, happiness, neutrality, sadness, and surprise.

Facial Expression Training Data The AffectNet [21] 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 96x96 pixels.

Natural Human Face Images for Emotion Recognition This unique dataset is curated from the Internet, encompassing more than 5,500 images manually labeled for eight emotional expressions. Each image captures real human expressions in grayscale format of 224x224 pixels.

3) Task Definition of Emotion Prediction with Four Situations: To assess ChatGPT-4's capacity for predicting emotional evolution, we performed a zero-shot prompting experiment. We curated a dataset of images spanning six emotions and provided the model with four unique situational prompts.

a) Prompt Engineering Strategy: Crucial to the reproducibility of Large Language Model research is the structure of the prompt. We utilized a zero-shot Chain-of-Thought (CoT) style prompt to encourage the model to reason about the facial features before predicting the emotional shift. The standard prompt template used is shown below:

This structured approach minimizes parsing errors and standardizes the output for automated scoring.

4) LLM querying, output parsing, and scoring pipeline: A practical challenge in LLM evaluation is that outputs are free-form by default. To enable automated scoring, we enforce a structured JSON output (Figure 1) and apply a strict parsing-and-normalisation pipeline:

TABLE III
LABEL HARMONISATION USED IN EXPERIMENTS AND MODELLING.

Source label	Model state E_i	Valence $g(\cdot)$
happiness / happy	Joy	$+1$
surprise	Surprise	$+1$ (often valence-ambiguous in practice)
neutral	(Neutral; evaluation-only)	0 (excluded from binary valence)
anger	Anger	-1
sadness / sad	Sadness	-1
disgust	Disgust	-1

System Prompt: You are an expert psychologist specializing in facial micro-expressions and emotional dynamics.
Input: [Image File]
User Query: 1. Identify the current emotion shown in the image. 2. Consider the following scenario: [Insert Scenario, e.g., "They receive a $1,000 reward"]. 3. Based on the initial emotion and the scenario, predict the most likely subsequent emotional state. 4. Provide a confidence score (1-3) for your prediction.
Output Format: JSON {current_emotion, predicted_emotion, confidence}

Figure 1. Zero-shot prompt template used for emotion prediction.

- **Output normalisation:** Map synonyms (e.g., "happy"→"happiness") and enforce the label set in Table III. If an output label is out-of-set, we map it to the nearest valence-consistent category when possible; otherwise it is marked as invalid.
- **Confidence as a score:** The confidence field (1–3) is treated as an ordinal score used for ROC/AUC where applicable. If confidence is missing, a default mid-score is assigned to avoid discarding samples.
- **Binary valence evaluation:** For valence-only tasks, Neutral is excluded and we map labels to $\{+, -\}$ via Table III.

Algorithm 2 summarises the end-to-end evaluation procedure used to produce confusion matrices and ROC/AUC.

Remark on undefined metrics (NaN). In some one-vs-rest settings, the denominator of Sensitivity ($TP + FN$) or Specificity ($TN + FP$) can be zero (e.g., if no samples of a target class remain after filtering, or if the model never predicts a class under a specific condition). In these cases the metric is mathematically undefined and we report NaN to avoid misleading values.

5) Preliminary Results: Table IV reports ChatGPT-4's predictions of emotion evolution. For images initially labeled negative, accuracy in negative contexts was 79.4%; in positive contexts it was 72.8%. For images initially labeled positive, accuracy was higher in positive than in negative contexts. This aligns with intuition: negative states are less likely to flip to positive under a positive context than to persist under a negative one; similarly, positive states are more stable in positive contexts.

Algorithm 2: Reproducible LLM evaluation pipeline.

```
Input: Dataset $\mathcal{D} = \{(x_n, y_n)\}_{n=1}^{N}$, prompt set $\mathcal{Q}$,
  label map $\phi(\cdot)$, valence map $g(\cdot)$.
Output: Confusion matrices;
  Accuracy/Sensitivity/Specificity; AUC where
  applicable.
foreach $(x_n, y_n) \in \mathcal{D}$ do
    foreach $q \in \mathcal{Q}$ do
        Query LLM with (image $x_n$, prompt $q$) $\rightarrow$ raw
          text $r$;
        Parse $r$ as JSON $\rightarrow$ $(\hat{y}^{\mathrm{cur}}, \hat{y}^{\mathrm{next}}, \hat{c})$;
        Normalise labels: $\hat{y} \leftarrow \phi(\hat{y})$;
        Update task-specific counters (six-way or
          valence-only);
        Store score $\hat{c}$ for ROC/AUC when defined;
    end
end
Compute metrics from confusion matrices; compute
  ROC/AUC from stored scores.
```

TABLE IV
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

Given safety considerations, we focus on anger, disgust, and sadness. For negative starting emotions followed by positive events (zero shot), the predictive precision ranks disgust, sadness, anger, with FPR of 78.3%, 71.7% and 68.3%, respectively. Anger appears most resistant to immediate improvement under positive events, whereas disgust—being semantically heterogeneous (e.g., dislike, contempt, displeasure)—shows the highest apparent accuracy.

6) Analysis and Discussion: Two issues emerged during evaluation. First, some dataset images diverge from common real-world interpretations. Second, there is a policy mismatch between ChatGPT-4's open-ended descriptions and the dataset's labeling guidelines: for example, an image tagged as "anger" in the dataset may be read as "sadness" or "confusion" by the model. These observations imply two practical paths. If strict adherence to the dataset taxonomy is not required, performance can be improved via prompt refinement (e.g., enumerating candidate emotions and contextual cues) and human-in-the-loop review. If strict adherence is required, prompt engineering alone is unlikely to suffice; supervised fine-tuning is the more appropriate strategy.

C. Emotion Prediction with Different Categories of Emotional Sentences

1) Dataset: In the second task, we added a dataset called MELD [22]. **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.

2) Task Definition: Part Two mirrors Part One by using the same image set, but augments each image with six emotion-conditioned utterances. To assess cross-model diversity, we run the identical protocol with the Doubao large language model [23] and compare outputs.

3) Preliminary Results: Overall accuracy (highest→lowest) is: happiness, surprise, neutral, anger, sadness, disgust. Within the "positive" set, happiness is generally most accurate; the main failure mode is a direct flip from happiness to anger, which yields the lowest accuracy for that class. Surprise and neutral track closely—consistent with ChatGPT-4's descriptions that treat both as valence-ambiguous. Among negative emotions, disgust is hardest to judge, reflected in the highest FPR (per the definition above) and the lowest accuracy. As in earlier tasks, zero-shot prompts are often insufficient for fine-grained negative labels: ChatGPT-4 reliably detects "negative" vs. "positive," but needs richer cues to distinguish specific negative categories.

The comparison model shows similar trends. Table VII contrasts accuracies for ChatGPT-4 and the Doubao LLM [23]. Doubao is notably less accurate on negative emotions, frequently defaulting to neutral or even (in zero-shot) misclassifying negatives as positive—patterns not observed with ChatGPT-4. While ChatGPT-4 may still confuse specific negative types (e.g., disgust vs. anger), it typically identifies that the affect is negative, explaining its stronger performance on emotion-evolution prediction.

Building on the earlier definitions, this section focuses on the Empirical ROC Area. The empirical Area Under the Curve (AUC) measures a model's ability to distinguish positives from negatives. From our data, sensitivities across the three datasets are broadly similar except for prompts expressing disgust. When the initial state varies, ChatGPT-4 finds disgust hardest to identify—e.g., in positive contexts it may reinterpret disgust as banter or a prank, reducing sensitivity. Specificity, however, is consistently strong, especially when the initial sentiment is positive, where predictions are nearly always correct. Taken together with the ROC curves, these results indicate that ChatGPT-4's emotion-conditioned sentence predictions perform better than anticipated.

IV. Discussion and Conclusion

A. Ethical Considerations and Limitations

While the ability of LLMs to predict emotional states offers significant benefits for empathetic human-computer interaction, it raises substantial ethical concerns. First, reliance on facial analysis for emotion detection has been criticized for potential bias; systems often perform poorly on underrepresented demographic groups if the training data is not diverse. In our study, although we used diverse datasets (Natural Human Faces), the underlying LLM's training distribution remains opaque.

Second, the "black box" nature of models like ChatGPT-4 presents a challenge for clinical deployment. If a model predicts a high risk of negative emotional spiraling (e.g., depressive states), the lack of explainability makes it difficult for human practitioners to trust the output without verification. Our Markov-based model attempts to mitigate this by imposing a mathematical structure on the output, but the core inference remains opaque.

Lastly, privacy is paramount. Real-time emotion tracking implies constant surveillance of user expressions. Any implementation of such systems must adhere to strict data privacy standards, ensuring that emotional data is processed locally where possible and not stored without explicit consent.

B. Failure Mode Taxonomy and Practical Implications

Across both tasks, errors are not uniformly distributed; they follow recurring patterns that are useful for both modelling and deployment.

(1) Valence-correct but category-wrong. A common outcome is that the model correctly predicts negative vs. positive affect while confusing specific negative labels (e.g., Disgust vs. Anger, or Disgust vs. Sadness). This directly supports the hierarchical assumption in Section II: a coarse valence layer can be stable even when fine-grained categorical boundaries are blurred.

(2) Ambiguity between Neutral and Surprise. Surprise is frequently treated as valence-ambiguous by the model, especially when facial cues are subtle. In practice, these

TABLE V
EXAMPLE OF SIX DIFFERENT CATEGORIES EMOTIONAL SENTENCES.

Dataset	Question 1	Question 2	Question 3	Question 4	Question 5	Question 6
	What is the emotion of this 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 is the emotion of this 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 is the emotion of this 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?

TABLE VI
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%	100.00%	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 VII
ACCURACY OF DIFFERENT LARGE LANGUAGE MODELS.

LLM	Negative Emotion Accuracy	Positive Emotion Accuracy
ChatGPT	68.89%	80.56%
Doubao	26.11%	40%

TABLE VIII
RESULT OF DATASET FOR SIX DIFFERENT CATEGORIES EMOTIONAL SENTENCES

Dataset	Parameter	Anger Sentence	Disgust Sentence	Happiness Sentence	Neutral Sentence	Sad Sentence	Surprise Sentence
Emotion Detection	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

confusions can inflate six-way errors while leaving valence-level performance relatively strong, depending on the mapping used.

(3) Dataset-label vs. commonsense mismatch. Several images in crowd-sourced datasets encode expression intensity, pose, or occlusion patterns that do not align cleanly with everyday interpretations. This produces "apparent errors" that may actually reflect label noise. In safety-sensitive settings, a conservative design choice is to prioritise reliable detection of *negative valence* over precise negative subtyping, and then escalate ambiguous cases to human review.

(4) Prompt sensitivity. The same image can yield different predicted transitions under small variations in wording, especially for negative emotions. This motivates the use of structured prompts (Figure 1), explicit candidate label sets, and (when feasible) repeated trials with aggregation to reduce variance.

C. Future Work

Our evaluation relies on static inputs (single images or texts), whereas real emotions evolve during interaction. Without real-time feedback to update predictions, immediate applicability to adaptive systems (e.g., conversational agents or monitoring tools) is limited. Although we center on ChatGPT-4 for image-based emotion recognition, future comparisons with other LLMs (e.g., Claude 3) and real-world trials are needed to assess robustness and generalizability. Improving transparency and accuracy may involve prompt refinement or supervised fine-tuning. Because responses are stochastic, single-trial outputs can vary; repeated runs with fixed seeds and averaged results would provide more reliable estimates and reduce variance-driven bias. Finally, judgments based solely on perceived emotional shifts can introduce labeling bias; careful protocol design and human review remain important.

We examined ChatGPT-4's zero-shot performance on image–text emotion interpretation and compared it with the Doubao model. ChatGPT-4 generally achieves higher accuracy, though it can confuse specific negative categories (e.g., classifying disgust as sadness/depressive affect). Targeted prompts and mental-health-aware guidance improve inference quality. Doubao underperforms ChatGPT-4 overall and, in zero-shot settings, more often maps negative affect to neutral or positive. For subjective tasks, we recommend prompt templates with explicit emotion taxonomies and illustrative exemplars; where strict adherence to dataset labels is required, supervised fine-tuning is likely necessary to align outputs with annotation guidelines. Finally, divergences between dataset tags and real-world perceptions can introduce bias; comparing human assessments with model outputs helps surface and correct such mismatches.

References

[1] Clement H. C. Leung and Zhifei Xu. "Predicting Emotion States Using Markov Chains". In: *BRAININFO 2025, The Tenth International Conference on Neuroscience and Cognitive Brain Information*. IARIA. 2025, pp. 7–16.

[2] Clement H. C. Leung, James J Deng, and Yuanxi Li. "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*. 2022, pp. 1–7.

[3] Ben Mann et al. "Language models are few-shot learners". In: *arXiv preprint arXiv:2005.14165* (2020).

[4] Long Ouyang et al. "Training language models to follow instructions with human feedback". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 27730–27744.

[5] *Open AI GPT 4*. https://openai.com/gpt-4. 2023.

[6] Tianlin Zhang et al. "Natural language processing applied to mental illness detection: a narrative review". In: *NPJ digital medicine* 5.1 (2022), p. 46.

[7] Davide Ciraolo et al. "Emotional Artificial Intelligence Enabled Facial Expression Recognition for Tele-Rehabilitation: A Preliminary Study". In: *2023 IEEE Symposium on Computers and Communications (ISCC)*. IEEE. 2023, pp. 1–6.

[8] *Fat cat incident*. https://sports.sohu.com/a/776021122_121856967.

[9] Russell Lewis and Joel Rose. *'I'm not okay,' off-duty Alaska pilot allegedly said before trying to cut the engines*. https://www.npr.org/2023/10/24/1208244311/alaska - airlines - off - duty - pilot - switch - off - engines. OCTOBER 25, 2023, 11:55 AM ET.

[10] AV Geetha et al. "Multimodal Emotion Recognition with deep learning: advancements, challenges, and future directions". In: *Information Fusion* 105 (2024), p. 102218.

[11] Rui Zhang et al. "Predicting emotion reactions for human–computer conversation: A variational approach". In: *IEEE Transactions on Human-Machine Systems* 51.4 (2021), pp. 279–287.

[12] Kailai Yang et al. "On the evaluations of chatgpt and emotion-enhanced prompting for mental health analysis". In: *arXiv preprint arXiv:2304.03347* (2023).

[13] Weixiang Zhao et al. "Is ChatGPT Equipped with Emotional Dialogue Capabilities?" In: *arXiv preprint arXiv:2304.09582* (2023).

[14] Hoai-Duy Le et al. "Multi-Label Multimodal Emotion Recognition With Transformer-Based Fusion and Emotion-Level Representation Learning". In: *IEEE Access* 11 (2023), pp. 14742–14751.

[15] Wei Zhang, Xuanyu He, and Weizhi Lu. "Exploring discriminative representations for image emotion recognition with CNNs". In: *IEEE Transactions on Multimedia* 22.2 (2019), pp. 515–523.

[16] Paul Ekman. *Facial expressions of emotion: New findings, new questions*. 1992.

[17] Plutchik Robert. *Emotion: Theory, research, and experience. vol. 1: Theories of emotion*. 1980.

[18] Ronak Kosti et al. "Emotion recognition in context". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 1667–1675.

[19] Ronak Kosti et al. "Context based emotion recognition using emotic dataset". In: *IEEE transactions on pattern analysis and machine intelligence* 42.11 (2019), pp. 2755–2766.

[20] Lutfiah Zahara et al. "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)*. IEEE. 2020, pp. 1–9.

[21] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. "Affectnet: A database for facial expression, valence, and arousal computing in the wild". In: *IEEE Transactions on Affective Computing* 10.1 (2017), pp. 18–31.

[22] Soujanya Poria et al. "Meld: A multimodal multi-party dataset for emotion recognition in conversations". In: *arXiv preprint arXiv:1810.02508* (2018).

[23] *Tik Tok*. https://www.doubao.com/chat/?channel=baidu_pz & source = db_baidu_pz_01 & keywordid = weizhi7. August 17, 2023.