MECoT: Markov Emotional Chain-of-Thought for Personality-Consistent Role-Playing

Anonymous ACL submission

Abstract

Large Language Models (LLMs) have shown remarkable capabilities in role-playing dialogues, yet they often struggle to maintain emotionally consistent and psychologically plausible character personalities. We present MECoT (Markov Emotional Chain-of-Thought), a framework that enhances LLMs' ability to generate authentic personality-driven dialogues through stochastic emotional transitions. Inspired by dual-process theory, MECoT combines a Markov-chain-driven emotional processor for intuitive responses with an LLM-based reasoning mechanism for rational regulation, mapped onto a 12-dimensional Emotion Circumplex Model. The framework dynamically adjusts emotional transitions using personality-weighted matrices and historical context, ensuring both emotional coherence and character consistency. We introduce the Role-playing And Personality Dialogue (RAPD) dataset, featuring diverse character interactions with fine-grained emotional annotations, along with novel metrics for evaluating emotional authenticity and personality alignment. Experimental results demonstrate MECoT's effectiveness, achieving 93.3% emotional accuracy on RAPD and substantially outperforming existing approaches. Our analysis reveals optimal emotional granularity (12-16 categories) and validates our data-driven personality optimization approach. Code and data are available at https://anonymous.4open.science/r/MECoT

1 Introduction

We think, fast and slow. — Kahneman (2011)

In recent years, Large Language Models (LLMs) have demonstrated remarkable capabilities in dialogue generation and emotion recognition. Despite significant progress in sentiment analysis (Zhang et al., 2023; Sun et al., 2023) and emotion generation (Lee et al., 2023; Li et al., 2024), current models exhibit fundamental limitations in role-playing scenarios where emotional authenticity and personality consistency are crucial. These limitations manifest in two critical ways: generating psychologically implausible emotional transitions and failing to maintain character-specific emotional patterns throughout extended interactions.

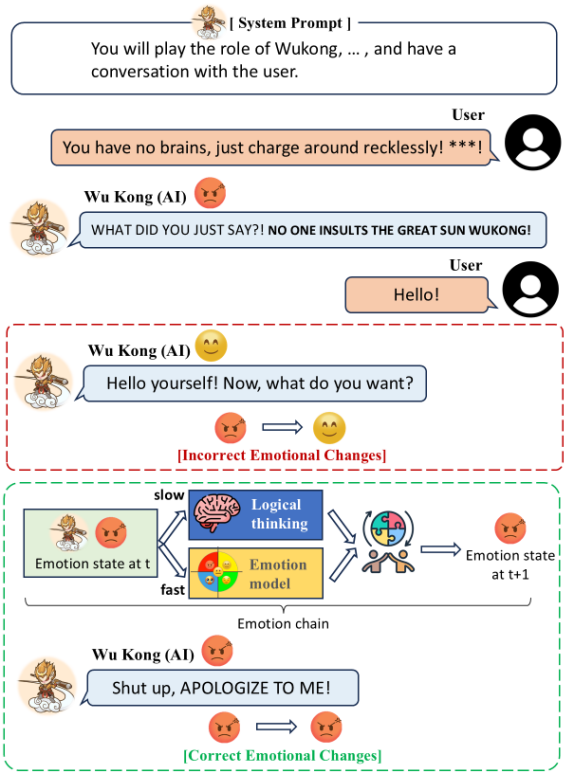

Figure 1: Example of emotional inconsistency in LLMs during role-playing dialogues. Wukong, characterized by a predisposition to anger and a tendency to maintain angry states, demonstrates the issue. The baseline model exhibits an abrupt transition from anger to happiness, whereas MECoT maintains appropriate emotional states through dual processes of instinct and reasoning, aligning with the character's established personality traits.

The challenge stems from the inherent complexity of human emotional processing, as articulated

in Kahneman's dual-process theory. Human emotional responses involve both rapid, intuitive reactions (System 1) and deliberate, rational regulation (System 2), modulated by individual personality traits and contextual factors. This nuanced interplay becomes evident in role-playing scenarios, where abrupt or inconsistent emotional transitions can significantly diminish user engagement and interaction quality. Figure 1 illustrates this issue through the character of Wukong, whose predisposition to anger requires careful emotional state management that current LLMs fail to provide.

Contemporary emotional modeling approaches suffer from two critical limitations. First, they often generate abrupt emotional transitions without sufficient contextual support, as illustrated in Figure 1 where models may shift suddenly from anger to happiness. These discontinuous transitions violate the principle of emotional gradualism and fail to reflect the regulatory role of character personality in emotional changes. Second, they treat each response independently, disregarding the cumulative effects and historical dependencies of emotional transitions. This becomes particularly problematic in extended dialogues where characters should exhibit consistent emotional patterns aligned with their established traits. For instance, an introverted and cautious character might suddenly display excessive extroversion and aggression, not only breaking user immersion but also violating fundamental principles of personality psychology. These limitations result in dialogue content that lacks long-term emotional coherence and fails to accurately reflect characters' emotional development trajectories through sustained interactions.

To address these challenges, we present **MECoT (Markov Emotional Chain-of-Thought)**, a framework that enhances LLMs' ability to generate authentic personality-driven dialogues through stochastic emotional transitions. MECoT implements a multi-level architecture that combines: 1) A bottom layer capturing basic emotional states through a 12-dimensional Emotion Circumplex Model, 2) A middle layer characterizing emotional transition probabilities via Markov chains, and 3) A top layer integrating character personality traits and historical context for emotional regulation. This design considers both immediate emotional stimuli and long-term factors such as character personality and dialogue history, ensuring the coherence and rationality of emotional changes. Through personality weight matrices and emotion adjustment

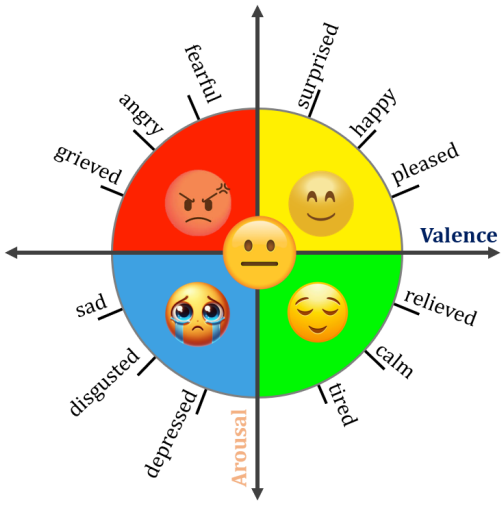

Figure 2: Emotion Circumplex Model with 12 basic emotions mapped in a two-dimensional space defined by valence and arousal.

mechanisms, MECoT can authentically reflect characters' emotional development trajectories while simulating both intuitive and rational analytical emotional responses based on Kahneman's dual-system. Our main contributions are:

1. We introduce an innovative framework MECoT that dynamically reconstructs personality-consistent emotional changes, significantly enhancing the coherence and authenticity of emotional modeling in role-playing dialogues.
2. We develop the Role-playing And Personality Dialogue (RAPD) dataset, featuring diverse character interactions with fine-grained emotional annotations, providing a robust benchmark for evaluating emotional dialogue generation.
3. We design a comprehensive evaluation metric system that assesses both emotional authenticity and personality consistency, enabling more nuanced analysis of model performance in role-playing scenarios.

2 Preliminaries

2.1 Emotion Circumplex Model

The Emotion Circumplex Model (Russell, 1980) is a fundamental theoretical framework in psychology for describing and quantifying emotional states. This model maps emotional states onto a two-dimensional plane as shown in Figure 2, representing different emotional states through two dimensions: Valence and Arousal.

The valence dimension represents the degree of positivity or negativity associated with an emotion, typically ranging from -1 to 1. Positive values correspond to positive emotions, such as happiness or satisfaction, while negative values correspond to negative emotions, such as sadness or anger. On the other hand, the arousal dimension captures the level of emotional activation, also within the range of -1 to 1. Higher arousal values signify highly activated emotional states, such as excitement or anger, whereas lower values indicate low activation states, such as calmness or fatigue.

In this model, any emotional state can be represented as a two-dimensional vector:

$$E_t = (v_t, a_t) \tag{1}$$

where v_t represents the valence value at time t, and a_t represents the arousal value at time t. For example, "excitement" might be represented as (0.8, 0.9), indicating high valence and high arousal, while "calm" might be represented as (0.3, -0.5), indicating moderate valence and low arousal. This quantitative representation enables us to precisely describe emotional states and provides a foundation for subsequent emotional transition modeling.

2.2 Markov Chains

A Markov chain is a probabilistic model that describes state transition processes, with its core characteristic being that the system's next state depends only on the current state, independent of historical states. In emotional modeling, Markov chains provide a natural framework for describing the evolution of emotional states (Cipresso et al., 2023).

Formally, a Markov chain consists of the following key elements:

1. State Space S: In emotional modeling, this is the set of all possible emotional states.
2. Transition Probability Matrix T: Matrix elements T_{ij} represent the probability of transitioning from state i to state j, satisfying:

$$T_{ij} \geq 0, \sum_j T_{ij} = 1. \tag{2}$$

3. Initial State Distribution: The probability distribution of the system's starting emotional state.

Markov chains offer unique advantages in emotional modeling: First, they naturally capture the gradual nature of emotional state changes, avoiding unreasonable jumps; second, by adjusting the transition probability matrix, we can easily incorporate character personality traits into the model; finally, the Markov property (that the next state depends only on the current state) aligns with the short-term dependency characteristics of human emotional changes, while through the introduction of additional weight matrices, we can also model longer-term emotional dependencies.

3 Methodology

3.1 Problem Formulation

In role-playing scenarios, our core objective is to achieve authentic and coherent emotional changes for characters. Formally, given the dialogue history H, character settings P, and the current emotional state E_t, we need to predict the next reasonable emotional state E_{t+1} and generate the corresponding dialogue response R. This process can be represented as:

$$f : (H, P, E_t) \rightarrow (E_{t+1}, R). \tag{3}$$

Here, the emotional state E is represented as a two-dimensional vector (v, a), where v and a denote valence and arousal, [need to explain these two] respectively. Our goal is to ensure that the generated emotional sequence $\{E_1, E_2, \ldots, E_t\}$ adheres to the natural principles of emotional change while reflecting the character's personality traits.

3.2 MECoT

MECoT is a dual-system framework for emotional transitions, inspired by fast and slow thinking in humans. It combines two processes: a Subconscious process (fast) and an Emotion reasoning process (slow), as illustrated in Figure 3(a). The Subconscious process uses a Markov chain based on the emotion circumplex model to simulate automatic emotional responses, influenced by emotional distance and character personality. The Emotion reasoning process uses large language models to perform multi-step reasoning for "rational" emotional responses aligned with the character. These two outputs jointly determine the next emotional state through sampling, guiding the system to generate dialogue that reflects natural emotional transitions while staying true to the character's personality.

3.2.1 Emotional State Representation

MECoT uses 12 basic emotions from the Emotion Circumplex Model as its discrete state space S:

$$S = \{e_i = (v_i, a_i) | i \in \{1, 2, \ldots, 12\}\} \tag{4}$$

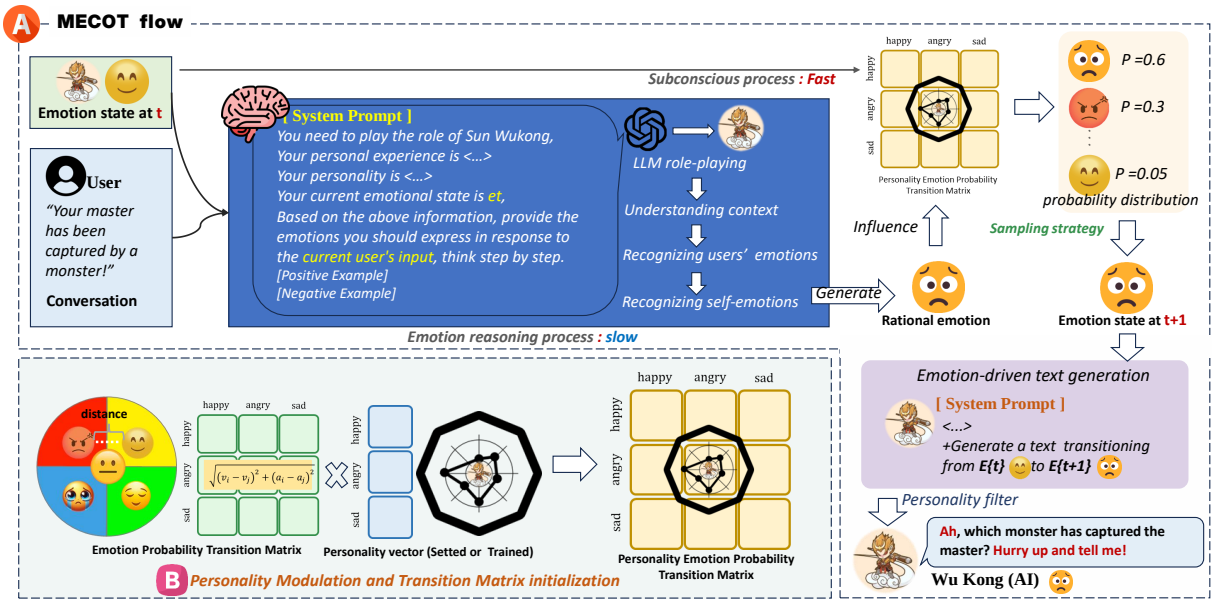

Figure 3: Overview of our proposed MECoT framework. (a) The process of MECOT generating emotional responses at time t. (b) The initialization and personality-based modulation of the emotional transition matrix.

where e_i denotes the i-th basic emotion, and v_i and a_i represent its valence and arousal values, respectively. The initial emotional transition matrix T_0 is constructed based on the Euclidean distance between emotional vectors as illustrated in Figure 3(b):

$$T_0[i,j] = \frac{\exp(-\|e_i - e_j\|_2)}{\sum_k \exp(-\|e_i - e_k\|_2)}. \quad (5)$$

Each row of the T_0 matrix represents the probability of transitioning from the current emotional state e_i to other emotional states e_j, while also reflecting the relative distances between emotion vectors.

3.2.2 Emotion Reasoning Process

To achieve deep emotional reasoning, we designed MECoT's emotional analysis process as a slow-thinking system. This system, based on LLMs, analyzes the impact of dialogue inputs on emotional states through character embodiment and multi-step reasoning. The reasoning process includes the following steps: first, the model needs to understand the character's personality traits and background experiences; second, it analyzes the current dialogue context and emotional state; finally, through analogy and reasoning, it weighs the rationality of different emotional responses to arrive at a "rational" emotional response that aligns with the character's traits.

This process can be formalized as a probability distribution:

$$\Delta E_{\text{input}} = (\delta v, \delta a) \sim \mathcal{N}(\mu, \Sigma) \quad (6)$$

where the expected value $\mu = (\mu_v, \mu_a)$ represents the optimal direction of emotional change derived through deep reasoning, and the covariance matrix Σ reflects the uncertainty in the reasoning process. The LLMs performs N independent sentiment inferences, with each inference outputting a rational emotion E^i_{input}, where $i = 1, 2, \ldots, N$. The corresponding change is obtained by taking the difference with E_t. The relevant prompts are provided in the Appendix Table 7.

3.2.3 Personality Modulation and Transition Matrix Update

To incorporate character personality traits into the model, we introduce a personality weight vector P:

$$P = \{p_j | j = 1, 2, \ldots, 12\} \quad (7)$$

where $p_j \in [0, 1]$ represents the degree of match between the j-th emotional state and the character's personality. The MECoT model initializes P using a dual-path scheme, with detailed information provided in the appendixA.3.

Considering the above factors, we update the transition matrix using the following formula:

$$T[i,j] = \Big(\underbrace{T_0[i,j]}_{\text{fast process}} + \underbrace{\beta \cdot \Delta E_{\text{input}} \cdot W}_{\text{slow process}} \Big) \cdot P[j]. \quad (8)$$

Here, W is a 2×12 weight matrix that maps the emotional change vector to the 12 basic emotions, and β is a coefficient that balances the initial transition probabilities with the influence of emotional

changes. The weight matrix $\mathbf{W} \in \mathbb{R}^{2\times n}$ is defined as:

$$\mathbf{W} = \begin{bmatrix} v_1 & v_2 & \cdots & v_n \\ a_1 & a_2 & \cdots & a_n \end{bmatrix} \tag{9}$$

where each column $\mathbf{W}[:, j] = [v_j, a_j]^T$ corresponds to the coordinates of a basic emotional state. When the emotion change ΔE_{input} aligns in direction with a certain emotionalstate E_j, the transition probability to that state is enhanced through dot product calculation.

3.2.4 Emotional State Selection Strategies

MECoT implements three distinct strategies for emotional state selection, each designed to handle different dialogue scenarios and character requirements.

Expected Value Strategy. The first strategy involves calculating a weighted average of potential emotional states, which is then mapped to the nearest basic emotion:

$$\bar{E}'_{t+1} = \sum_{j=1}^{12} \mathbf{p}_j \cdot e_j, \quad E_{t+1} = \arg\min_{e_k} \|\bar{E}'_{t+1} - e_k\|_2. \tag{10}$$

This approach is particularly effective in daily conversations, where smooth emotional transitions are essential.

Maximum Probability Strategy. The second strategy focuses on directly selecting the emotional state with the highest probability:

$$E_{t+1} = e_j, \text{ where } j = \arg\max_m(\mathbf{p}_m). \tag{11}$$

This method shines during critical plot moments that demand clear and decisive emotional shifts.

Probabilistic Sampling Strategy. The third strategy introduces an element of controlled randomness through threshold-based sampling:

$$\mathbf{P}(E_{t+1} = e_j) \propto T[i, j] \cdot \mathbb{1}(T[i, j] > \theta). \tag{12}$$

While enhancing the variety of character responses, this approach ensures that emotional coherence is still maintained.

3.2.5 Emotion-Driven Text Generation

MECoT employs a hierarchical generation strategy, first generating response content based on emotional state transitions and character settings, then modulating the generated content through a Personality Filter. The **Personality Filter** contains two key components: **Character Modulator** and **Style Transformer**. Formally, the generation process can be represented as:

$$R = F_s(F_c(M(H, P, E_t, E_{t+1}))) \tag{13}$$

where M is a large language model, F_c is the character modulation function, and F_s is the language style transformation function.

3.2.6 Personality Weight Optimization

To ensure that the personality weights can dynamically adapt to specific scenarios, MECoT introduces a heuristic optimization mechanism. When the predicted emotion E^{pred}_{t+1} approaches the true emotion E^{true}_{t+1}, the model updates weights through:

$$P[j_{\text{true}}] \leftarrow P[j_{\text{true}}] + \alpha \cdot (1 - P[j_{\text{true}}]), \tag{14}$$

$$P[j] \leftarrow P[j] - \alpha \cdot P[j] \quad (\forall j \neq j_{\text{true}}). \tag{15}$$

To ensure optimization stability, the model sets a weight lower bound $P[j] \geq \epsilon$ (e.g., $\epsilon = 0.05$), introduces a momentum term $P_t = \beta P_{t-1} + (1 - \beta)P_{\text{new}}$ to avoid weight oscillation, and maintains the sum of weights at 1. This data-driven optimization approach enables MECoT to continuously improve emotional transition accuracy while maintaining character personality consistency.

3.3 Evaluation Method

To comprehensively evaluate the performance of the MECoT model, we designed evaluation metrics across three dimensions: emotional transition accuracy, emotional change trends, and character personality consistency.

For emotional transition accuracy, we employ two metrics: **Emotional Classification Accuracy** and **Emotional Distance**. The emotional classification accuracy measures the degree of match between the emotions in generated text and target emotions, formally defined as:

$$Acc = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}(C(x_i) = y_i)$$

where N is the total number of test samples, $C(x_i)$ represents the classification result of the emotion classifier on generated text x_i, y_i is the target emotion category, and $\mathbb{1}(\cdot)$ is the indicator function. The emotional distance measures the Euclidean distance between generated and target emotions in the continuous valence-arousal space:

$$ED = \frac{1}{N} \sum_{t=1}^{N} \sqrt{(v_t - v_t^{gt})^2 + (a_t - a_t^{gt})^2} \tag{16}$$

		DailyDialog	**RAPD (ours)**
Num. Roles		✘	**73**
Num. Dialogues		13K	8.5K
Avg. Turn per Dialogue		7.9	**13.3**
Avg. Length per Dialogue		52.3	**87.1**
Num. Emotion type		7	**12**
Avg. Emotion per Dialogue		2.4	**6.7**
Type	low (CED<1)	8.5k	5.2k
	medium (1<CED<3)	3.2k	2.8k
	high (3<CED)	0.3k	**0.5k**

Table 1: Comparison between DailyDialog and RAPD.

where (v_t, a_t) and (v_t^{gt}, a_t^{gt}) represent the coordinates of generated and ground truth emotions in the valence-arousal space at time t, respectively.

To evaluate the coherence of emotional changes, we introduce the **Emotional Trend Correlation Coefficient** (ETCC). This metric combines the Pearson correlation coefficients for both valence and arousal sequences:

$$r_v = \text{corr}(v, v^{gt}), r_a = \text{corr}(a, a^{gt}) \quad (17)$$

The final ETCC is calculated as $ETCC = \sqrt{\frac{{r_v}^2 + {r_a}^2}{2}}$, with higher values indicating better capture of emotional change trends.

For evaluating character **Personality Consistency**, we adopt a LLM-based (deepseek-r1) evaluation method (Ahn et al., 2024). This metric assesses whether the generated content aligns with the character's thought patterns, speaking style, tone, emotional reactions, and behavioral patterns.

4 Experiments and Analysis

4.1 Dataset and Baseline Methods

Dataset. We first evaluate general conversational abilities on the DailyDialog (Li et al., 2017), which does not rely on specific character traits. For character-based dialogue testing, we introduce RAPD (Role-specific Affective Persona Dialog), the first dataset specifically designed for emotional dialogues grounded in distinct character personas. Compared to DailyDialog, RAPD offers significant advancements, featuring a richer and more nuanced structure with a larger variety of characters, a wider range of emotion types, and a higher frequency of emotional transitions within dialogues. Furthermore, RAPD categorizes dialogues based on Cumulative Emotional Distance (CED) into low, medium, and high emotional variation, making it a robust resource for studying complex emotional dynamics in character-driven interactions.

Baseline Methods. We evaluate MECoT against 5 baseline methods:

	Methods	$Acc(\%) \uparrow$	$ED \downarrow$	$ETCC \uparrow$
	deepseek-v3 (*0-shot*)	74.8	0.31	0.57
Self	+ 2-shot	86.5	0.24	0.68
	+ CoT	88.2	0.21	0.73
	+ ECoT	89.6	0.19	0.77
	+ MECoT (ours, Default)	**93.4**	**0.13**	**0.91**

Table 2: Performance Comparison of MECoT and Baselines on the DailyDialog Dataset. Demonstrate emotional abilities unrelated to the role.

1. Zero-shot: Direct use of pre-trained models.
2. Chain-of-Thought (Zhang et al., 2024a): Step-by-step reasoning through prompts.
3. Emotional Chain-of-Thought (ECoT) (Li et al., 2024): Emotional reasoning with CoT.
4. RAG+ECoT: Combining retrieval-augmented generation (Lewis et al., 2020) with ECoT.
5. Finetuned: Fine-tuning pre-trained models using role-playing data to enhance their ability to dialogue understanding (Hu et al., 2021).

To adapt to our current task, we made appropriate modifications to the baseline prompt. For our proposed MECoT method, we designed three configurations: a default setup with equal weights for all P components (Default), a manually adjusted setup based on character impressions (Setted), and a trained setup where parameters are optimized acroding to 3.2.6 using the Setted configuration as a baseline (Trained).

4.2 Experimental Results Analysis

4.2.1 Baseline Performance Evaluation

We first validated MECoT's effectiveness on the DailyDialog dataset. As shown in Table 2, compared to zero-shot (74.8%), MECoT with default parameters improved accuracy by 18.6 percentage points (93.4%), reduced emotional distance by 0.18, and increased ETCC by 0.34. This result validates the superiority of the dual-system architecture in general dialogue (character-independent) scenarios, particularly excelling in handling progressive emotional changes in daily conversations.

Having ensured performance on character-independent datasets, we further explored performance in character-specific scenarios. On the RAPD dataset, our MECoT method performed excellently across all three models, particularly after parameter optimization (Trained), significantly improving emotional classification accuracy, emotional distance, and personality consistency metrics (see Table 3). This result validates MECoT's dual-system architecture's adaptability and advantages in handling complex emotional scenarios, while

Methods	Acc (%) ↑				ED ↓	Personality Consistency (1-5) ↑	ETCC ↑
	low	medium	high	Avg.			
Deepseek-chat (`deepseek-v3-671B`)							
zero-shot	86.7±2.8	73.6±3.5	64.0±4.9	74.8	0.37±0.08	3.6±0.5	0.54
CoT	88.5±2.5	76.0±3.2	68.0±4.5	77.5	0.34±0.07	3.8±0.4	0.58
ECoT	89.0±2.3	77.5±3.0	69.5±4.2	78.7	0.32±0.06	3.9±0.3	0.60
RAG + ECoT	91.0±2.1	81.0±2.8	73.0±3.8	81.7	0.28±0.05	4.0±0.3	0.72
MECoT (ours, Default)	89.5±2.2	78.0±2.9	70.0±4.0	79.2	0.30±0.06	4.0±0.3	0.68
MECoT (ours, Setted)	92.0±1.9	83.0±2.5	81.0±3.5	84.3	0.18±0.04	4.3±0.3	0.83
MECoT (ours, Trained)	**96.5±1.7**	**91.0±2.2**	**86.0±3.2**	**90.2**	**0.09±0.03**	**4.5±0.2**	**0.90**
Meta-Llama-3-70b							
zero-shot	78.4±4.6	64.2±4.0	58.2±4.2	66.9	0.43±0.10	3.2±0.6	0.47
CoT	80.0±4.2	67.0±3.8	60.0±3.9	69.0	0.40±0.09	3.4±0.5	0.50
ECoT	81.5±3.9	68.5±3.5	61.5±3.7	70.5	0.38±0.08	3.5±0.4	0.52
RAG + ECoT	85.0±3.5	72.0±3.2	67.5±3.5	74.0	0.33±0.07	3.8±0.4	0.65
Finetuned	82.0±3.8	69.0±3.6	64.0±3.8	71.7	0.34±0.07	**4.4±0.4**	0.58
MECoT (ours, Setted)	86.0±3.3	78.0±3.0	74.0±3.3	78.7	0.28±0.06	4.1±0.3	0.75
MECoT (ours, Trained)	**94.5±3.0**	**84.0±2.8**	**81.4±3.0**	**84.5**	**0.17±0.05**	4.3±0.3	**0.82**
Deepseek-reasoner (`deepseek-r1-671B`)							
zero-shot	95.8±1.4	84.3±2.2	82.4±3.0	87.5	0.16±0.04	4.3±0.3	0.84
CoT	96.0±1.3	85.0±2.0	83.0±2.8	88.0	0.15±0.03	4.3±0.3	0.85
ECoT	96.2±1.2	85.5±1.9	83.5±2.7	88.4	0.14±0.03	4.4±0.2	0.86
RAG + ECoT	97.5±0.8	87.0±1.7	85.0±2.5	89.7	0.12±0.02	4.5±0.2	0.90
MECoT (ours, Default)	96.5±1.1	86.0±1.8	84.0±2.6	88.8	0.13±0.03	4.4±0.2	0.87
MECoT (ours, Setted)	97.5±0.9	88.0±1.5	86.0±2.3	90.5	0.10±0.02	4.6±0.2	0.92
MECoT (ours, Trained)	**98.0±0.8**	**92.0±1.3**	**90.5±2.0**	**93.3**	**0.06±0.01**	**4.7±0.2**	**0.94**

Table 3: Comparative Performance of MECoT and Baselines Across Different Models and Settings

demonstrating that data-driven optimization strategies can further enhance the model's ability to capture character-specific emotions.

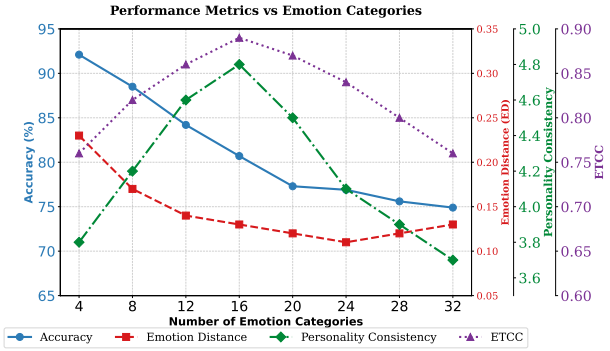

Figure 4: Impact of Emotion Category Granularity on Model Performance

Furthermore, we compared MECoT with Llama3-70b, fine-tuned for role-playing tasks, and observed that while it showed improvements in character consistency, its performance in emotion-chain-related dialogue abilities lagged significantly behind MECoT. On the other hand, the Deepseek-r1 model demonstrated the best performance across all scenarios, achieving an impressive 87.5% accuracy even in zero-shot settings. This underscores the advantages of models specifically designed for reasoning tasks in emotional understanding and generation. These findings highlight the critical importance of strong reasoning capabilities for managing complex emotional transitions and point to a promising direction: leveraging reinforcement learning with emotion-chain data to enhance LLMs.

4.2.2 Emotional Granularity Study

We investigated the impact of the number of emotion categories (from 4 to 32) in the emotion circumplex model on model performance, as shown in the Figure 4. The experiments revealed that as the number of emotion categories increased, emotional classification accuracy gradually decreased from 92.1% to 70.5%, reflecting the increased difficulty of classification with finer-grained emotion divisions. However, emotional distance reached its minimum value (0.11) at 16 emotion categories before slightly rebounding, indicating that finer-grained emotion divisions help generate responses closer to target emotions. Personality consistency

Dataset	Method	Acc (%) ↑	ED ↓	Consistency ↑
	MECoT (setted)	73.5	0.42	3.8
Harry(100)	MECoT (trained)	76.0	0.38	4.1
Harry(500)	MECoT (trained)	88.9	0.18	4.3
Harry(1k)	MECoT (trained)	**96.4**	**0.07**	**4.6**

Table 4: Performance comparison between MECoT (setted) and MECoT (trained) on the Harry Potter dataset.

peaked (4.8) at 16 emotion categories before gradually declining, suggesting that too many emotion categories may weaken the model's ability to stably simulate character personalities. Meanwhile, ETCC reached its highest value at 16 emotion categories before gradually declining, demonstrating that moderate emotional granularity helps capture emotional change trends. Therefore, experimental results indicate that optimal performance balance is achieved with 12 to 16 emotion categories.

4.2.3 Personalization Training Analysis

Using Harry Potter characters as a case study, we investigated the impact of training data volume on model performance. Our experiments revealed that with only 100 training samples, performance improvements were quite limited due to imbalanced emotion category distribution. However, when training data was increased to 1,000 samples, we observed significant performance gains (see Table 4). This demonstrates that with sufficient character-specific data, MECOT can effectively learn and simulate the unique emotional expression patterns of characters.

This approach is particularly advantageous for characters with substantial appearances in source material, while characters with fewer appearances require manual adjustment of transition matrix P to match our conceptual impressions. Interestingly, we found that social media dialogues can be collected to shape virtual representations of real individuals, as these samples are typically abundant. The appendix demonstrates this "data-to-personality" reverse engineering process.

5 Related Work

5.1 LLMs Role-Playing

LLM-based agents have shown advanced abilities such as planning, reflection, and tool use (Yao et al., 2024, 2022; Shinn et al., 2024). A key method is **role-playing**, where personas embedded into prompts enable models to simulate traits and behaviors, adapting flexibly to diverse scenarios.

In multi-agent settings (Guo et al., 2024; Liu et al., 2023), role-playing supports collaboration on complex tasks. Park et al. (Park et al., 2023) introduced generative agents mimicking human behavior, while Shao et al. (Shao et al., 2023) developed Character-LLM to simulate historical figures, demonstrating strong role memory.

Kong et al. (Kong et al., 2024) proposed **self-prompt tuning**, training LLMs to autonomously generate role prompts, outperforming traditional methods. Carlander et al. (Carlander et al., 2024) explored tabletop role-playing with Controlled Chain of Thought (CCoT) for context-based reasoning. Shen et al. (Shen et al., 2024) simulated 16 Myers-Briggs personality types to evaluate adaptability and decision-making.

5.2 Chain-of-Though in Affective generation

Affective generation (AG) integrates emotion and reasoning to produce emotionally rich responses. Cue-CoT (Wang et al., 2023) infers user states from linguistic cues, while MT-ISA (Lai et al., 2024) improves emotion recognition through multi-task learning.

CoT reasoning enhances emotional intelligence. ECoT (Zhang et al., 2024c) uses emotion-aware prompting, evaluated via the Emotional Generation Score (EGS). EBG (Zhu et al., 2024) performs emotional reasoning before response generation, improving empathy. DSC (Chen and Liu, 2023) dynamically generates counseling strategies for mental health tasks. The COOPER dialogue framework (Cheng et al., 2024) coordinates multiple pecialized agents, each focusing on specific aspects of dialogue goals, to generate emotional responses.

Strategic planning is key in complex emotional tasks. ProCoT (Deng et al., 2023) enables goal-driven responses, while ECoT and ES-CoT (Zhang et al., 2024b) improve emotional consistency through recognition and regulation. Chen et al. (Chen et al., 2024) proposed causal-driven empathy generation using external knowledge (e.g., COMET) to enhance reasoning and diversity.

6 Conclusion

MECoT integrates Markov chains with LLMs to achieve authentic emotional transitions in character-based dialogues. The study highlights the importance of optimal emotional granularity and data-driven personality optimization for model performance.

7 Limitations

Despite its strong performance, MECoT has limitations that warrant further exploration. First, its cross-cultural adaptability remains a challenge, as emotional expressions and transitions can vary significantly across cultures, which may affect its generalization in diverse settings. Second, the framework currently focuses on text-based interactions, limiting its applicability in real-time, multimodal scenarios where audio, visual, and contextual cues play a crucial role in emotion recognition and response generation.

In addition, the reliance on pre-defined emotional categories and a structured personality model may constrain its flexibility in highly dynamic or open-domain situations. Future work should aim to address these limitations by integrating multimodal emotion processing, improving adaptability to diverse cultural contexts, and refining the framework for real-time, interactive applications.

8 Ethics Statement

This work focuses on advancing emotionally intelligent dialogue systems, and we acknowledge the ethical implications associated with such technologies. While MECoT is designed to enhance engagement and realism in role-based dialogues, misuse of this technology could lead to manipulative or deceptive interactions, especially in sensitive or vulnerable contexts. To mitigate these risks, we emphasize transparency in the system's purpose and usage, ensuring users are aware when interacting with an AI model.

Additionally, the dataset and model development were conducted with ethical considerations in mind, avoiding the use of harmful or biased content. However, we recognize that emotional modeling, especially in diverse cultural or social contexts, may inadvertently reinforce stereotypes or biases. Future work will prioritize fairness and inclusivity, alongside mechanisms to detect and address potential ethical concerns in deployment.

References

Jaewoo Ahn, Taehyun Lee, Junyoung Lim, Jin-Hwa Kim, Sangdoo Yun, Hwaran Lee, and Gunhee Kim. 2024. Timechara: Evaluating point-in-time character hallucination of role-playing large language models. *arXiv preprint arXiv:2405.18027*.

Deborah Carlander, Kiyoshiro Okada, Henrik Engström, and Shuichi Kurabayashi. 2024. Controlled chain of thought: Eliciting role-play understanding in llm through prompts. In *2024 IEEE Conference on Games (CoG)*, pages 1–4. IEEE.

Qi Chen and Dexi Liu. 2023. Dynamic strategy chain: Dynamic zero-shot cot for long mental health support generation. *arXiv preprint arXiv:2308.10444*.

Xinhao Chen, Chong Yang, Man Lan, Li Cai, Yang Chen, Tu Hu, Xinlin Zhuang, and Aimin Zhou. 2024. Cause-aware empathetic response generation via chain-of-thought fine-tuning. *arXiv preprint arXiv:2408.11599*.

Yi Cheng, Wenge Liu, Jian Wang, Chak Tou Leong, Yi Ouyang, Wenjie Li, Xian Wu, and Yefeng Zheng. 2024. Cooper: Coordinating specialized agents towards a complex dialogue goal. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17853–17861.

Pietro Cipresso, Francesca Borghesi, and Alice Chirico. 2023. Affects affect affects: A markov chain. *Frontiers in Psychology*, 14:1162655.

Yang Deng, Lizi Liao, Liang Chen, Hongru Wang, Wenqiang Lei, and Tat-Seng Chua. 2023. Prompting and evaluating large language models for proactive dialogues: Clarification, target-guided, and non-collaboration. *arXiv preprint arXiv:2305.13626*.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. Large language model based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Daniel Kahneman. 2011. Thinking, fast and slow. *Farrar, Straus and Giroux*.

Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, Xin Zhou, Jiaming Zhou, and Haoqin Sun. 2024. Self-prompt tuning: Enable autonomous role-playing in llms. *arXiv preprint arXiv:2407.08995*.

Wenna Lai, Haoran Xie, Guandong Xu, and Qing Li. 2024. Multi-task learning with llms for implicit sentiment analysis: Data-level and task-level automatic weight learning. *arXiv preprint arXiv:2412.09046*.

Yoon Kyung Lee, Inju Lee, Minjung Shin, Seoyeon Bae, and Sowon Hahn. 2023. Chain of empathy: Enhancing empathetic response of large language models based on psychotherapy models. *arXiv preprint arXiv:2311.04915*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation

for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Cheng Li, Ziang Leng, Chenxi Yan, Junyi Shen, Hao Wang, Weishi Mi, Yaying Fei, Xiaoyang Feng, Song Yan, HaoSheng Wang, et al. 2023. Chatharuhi: Reviving anime character in reality via large language model. *arXiv preprint arXiv:2308.09597*.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. *arXiv preprint arXiv:1710.03957*.

Zaijing Li, Gongwei Chen, Rui Shao, Dongmei Jiang, and Liqiang Nie. 2024. Enhancing the emotional generation capability of large language models via emotional chain-of-thought. *arXiv preprint arXiv:2401.06836*.

Ruibo Liu, Ruixin Yang, Chenyan Jia, Ge Zhang, Denny Zhou, Andrew M Dai, Diyi Yang, and Soroush Vosoughi. 2023. Training socially aligned language models on simulated social interactions. *arXiv preprint arXiv:2305.16960*.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22.

James A Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-llm: A trainable agent for role-playing. *arXiv preprint arXiv:2310.10158*.

Chenglei Shen, Guofu Xie, Xiao Zhang, and Jun Xu. 2024. On the decision-making abilities in role-playing using large language models. *arXiv preprint arXiv:2402.18807*.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2024. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36.

Xiaofei Sun, Xiaoya Li, Shengyu Zhang, Shuhe Wang, Fei Wu, Jiwei Li, Tianwei Zhang, and Guoyin Wang. 2023. Sentiment analysis through llm negotiations. *arXiv preprint arXiv:2311.01876*.

Quan Tu, Shilong Fan, Zihang Tian, and Rui Yan. 2024. Charactereval: A chinese benchmark for role-playing conversational agent evaluation. *arXiv preprint arXiv:2401.01275*.

Hongru Wang, Rui Wang, Fei Mi, Yang Deng, Zezhong Wang, Bin Liang, Ruifeng Xu, and Kam-Fai Wong. 2023. Cue-cot: Chain-of-thought prompting for responding to in-depth dialogue questions with llms. *arXiv preprint arXiv:2305.11792*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2024. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.

Linhao Zhang, Li Jin, Guangluan Xu, Xiaoyu Li, and Xian Sun. 2024a. Cot: A generative approach for hate speech counter-narratives via contrastive optimal transport. *arXiv preprint arXiv:2406.12304*.

Tenggan Zhang, Xinjie Zhang, Jinming Zhao, Li Zhou, and Qin Jin. 2024b. Escot: Towards interpretable emotional support dialogue systems. *arXiv preprint arXiv:2406.10960*.

Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Jialin Pan, and Lidong Bing. 2023. Sentiment analysis in the era of large language models: A reality check. *arXiv preprint arXiv:2305.15005*.

Yiqun Zhang, Xiaocui Yang, Xingle Xu, Zeran Gao, Yijie Huang, Shiyi Mu, Shi Feng, Daling Wang, Yifei Zhang, Kaisong Song, et al. 2024c. Affective computing in the era of large language models: A survey from the nlp perspective. *arXiv preprint arXiv:2408.04638*.

Jiahao Zhu, Zijian Jiang, Boyu Zhou, Jionglong Su, Jiaming Zhang, and Zhihao Li. 2024. Empathizing before generation: A double-layered framework for emotional support llm. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pages 490–503. Springer.

A Appendix

A.1 Dataset construction

A.1.1 Data Source

Our dataset construction integrated public datasets with manual annotation to ensure diversity and high quality. The specific steps are as follows:

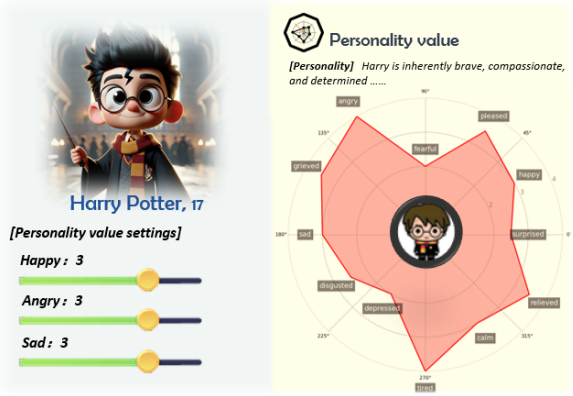

Figure 5: Illustration of Personality Weights for Harry Potter (Setted)

1. **Public Dataset Selection**
   We selected multiple public dialogue datasets as foundational corpus sources, for example:
   - **CharacterEval (Tu et al., 2024)**: Provides dialogue data with character personality settings across multiple scenarios.
   - **ChatHaruhi (Li et al., 2023)**: Provides Harry Potter novel multi-scenario dialogue data.

   These datasets laid a solid foundation for dialogue corpus, covering diverse conversation topics.

2. **Manual Expansion**
   To address insufficient dataset quantity, we also crawled data from movie and TV script websites [1]. We recommend including as many dialogues per character as possible to ensure coverage across emotional categories. To compensate for the lack of certain emotion categories or personality traits in public datasets, we manually designed specific scenarios (such as conflict resolution, celebratory events) to expand rare emotion categories.

3. **Data Cleaning and Preprocessing**
   Data cleaning steps included:
   - **Repetition and Noise Filtering**: Removing duplicate dialogues and meaningless noise data.
   - **Dialogue Segmentation**: Dividing long dialogues into multiple turns while preserving context information for each turn.
   - **Character Annotation**: Clearly labeling participating characters and their personality traits for each dialogue turn.
   - **Removing Texts with Unclear Emotions**: Eliminating content where emotions are not evident.
   - **Removing Biased Content**: Eliminating content that might involve cultural, gender, or other sensitive issues to ensure fairness and diversity in the dataset.

[1] https://subslikescript.com/

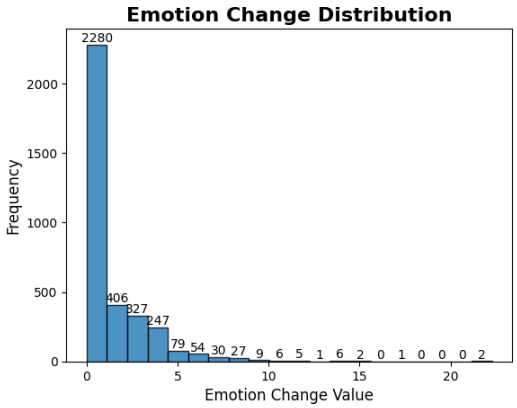

Figure 6: Distribution of Emotional Changes in Chinese Datasets (CED)

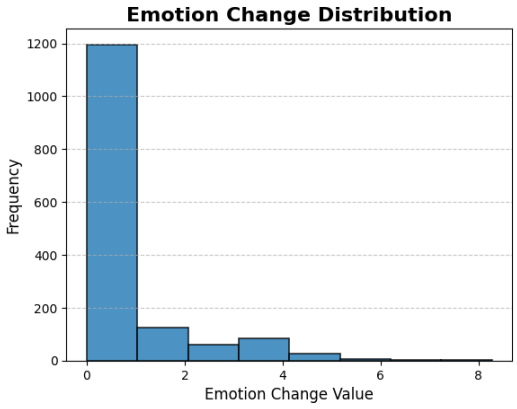

Figure 7: Distribution of Emotional Changes in Chinese Datasets (CED)

A.1.2 Emotion Annotation

Emotion annotation is a crucial component of this dataset, aiming to accurately label emotional states

for each dialogue turn. The specific process is as follows:

1. **Emotion Category Definition**
   We categorized emotional states into 12 emotion types, including **anger**, **joy**, **sadness**, **fear**, etc.

2. **Annotation Method**

   - **Automatic Annotation**: For texts with obvious emotions,LLMs were used for automatic annotation (Zhang et al., 2023; Sun et al., 2023), followed by human verification.

3. **Annotation Guidelines**
   To ensure annotation quality, we developed detailed annotation protocols, including:
   - Definitions and examples for each emotion category.
   - Guidelines for determining emotion valence and arousal labels based on context.
   - Priority rules for ambiguous multi-emotion scenarios (e.g., selecting the most prominent emotion when multiple emotions overlap).

4. **Sample Annotation**
   Example annotation as follows:
   - Dialogue content:
     A: "I finally got that promotion!"
     B: "That's amazing! Congratulations!"
     – A: Emotion category: **joy**
     – B: Emotion category: **excitement**

A.1.3 Dataset statistics

Figures 6 and 7 show the basic distribution of emotional changes (CED) in the Chinese and English datasets.

A.2 Distribution of Dataset and Its Impact on Model Performance

A.2.1 Dataset Partitioning

The datasets used in our experiments were partitioned as follows:

- **DailyDialog Dataset**: Used to evaluate general emotional dialogue capabilities independent of specific character features. The dataset was split into training, validation, and testing sets in an 8:1:1 ratio.
- **RAPD Dataset**: Specifically designed for role-playing emotional dialogue tasks, encompassing 73 characters, 12 emotion types, and 8.5K dialogues, with high emotional fluctuation scenarios accounting for 5.9%.
- **Data Balancing**: To mitigate biases caused by class imbalance, we applied SMOTE oversampling to rare emotion categories and undersampling to high-frequency emotion categories.

A.2.2 Emotion Category Distribution and Model Performance

Analysis of accuracy rates and emotional distances across different emotion categories revealed:

- Lower accuracy rates for anger and sadness categories, primarily due to high semantic overlap with adjacent emotion categories (such as fear and disappointment).
- In the RAPD dataset, scenarios with high emotional fluctuation showed lower ETCC scores, indicating that complex emotional transitions pose greater challenges to the emotion reasoning module.

A.2.3 Impact of Training Data Volume

Experiments on the Harry Potter dataset demonstrated:

1. When training samples increased from 100 to 1000:
   - Emotion classification accuracy improved from 73.5% to 96.4%
   - Emotional Distance (ED) decreased from 0.42 to 0.07

   These results indicate that sufficient character-specific data significantly enhances emotional consistency and personalized simulation effectiveness.

2. In scenarios with limited training data:
   - Increased randomness in the emotion transition matrix led to greater emotional fluctuations
   - Future work could address data insufficiency through data augmentation or transfer learning techniques

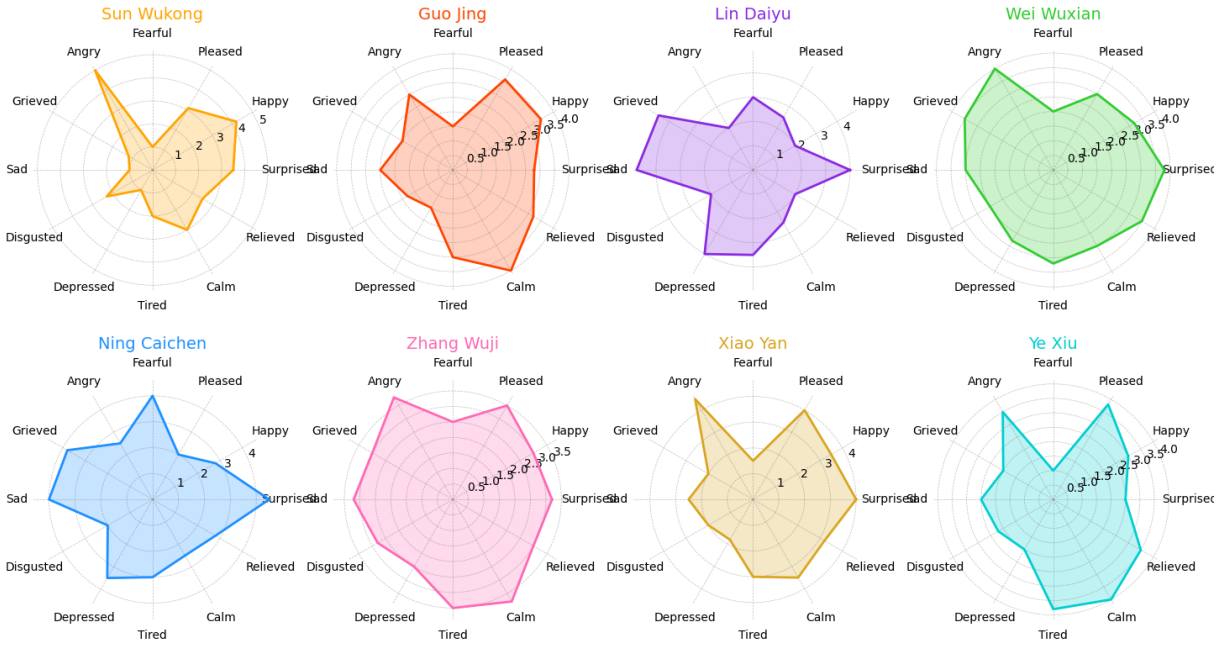

Figure 8: Visualization of the personalities of eight characters 1

Figure 9: Visualization of the personalities of eight characters 2

A.3 Personality Weight Design

To set personality weights (or vectors) for a character, we can utilize two easy approaches: **manual design** and **generation via LLMs**.

A.3.1 Manual Design

Personality weights can be manually assigned based on predefined impressions or character archetypes, as illustrated in Figure 5. For example, using the Big Five personality traits (Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism), a highly extroverted and agreeable character might be assigned weights such as $[0.8, 0.6, 0.9, 0.7, 0.3]$. These weights can then be normalized and integrated into the model to influence emotional transitions and responses.

A.3.2 LLM-Generated Weights

Large language models can be prompted to generate personality traits or weights based on descriptive input. By providing a detailed prompt about the character's background, preferences, and behavior, the model can output a structured personality profile or weights. For example, a prompt might describe the character as "an empathetic and optimistic individual who is highly energetic but prone to occasional impulsiveness," and the model can generate corresponding weights for predefined personality dimensions.

A.3.3 Example Prompt for LLM-Generated Personality Weights

Prompt Template:
You are designing a fictional character for a role-playing dialogue system. This character should have a well-defined personality based on the following description:

- **Name:** Alex
- **Description:** Alex is cheerful and outgoing, loves meeting new people, and focuses on positivity. However, they can sometimes ignore risks due to excessive optimism.
- **Personality Dimensions:** Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism.

Assign a numerical weight (0 to 1) to each dimension, where 1 indicates a strong trait and 0 an absent trait. Provide weights and brief explanations.
Expected Output:
Personality weights for Alex:

- **Openness:** 0.85 (Alex is curious and enjoys exploring new ideas.)
- **Conscientiousness:** 0.65 (Organized but occasionally overlooks details.)
- **Extraversion:** 0.95 (Highly sociable and outgoing.)
- **...**

These weights can then be normalized or scaled as needed and integrated into the model to adjust emotional dynamics and dialogue coherence based on the character's personality. The personality weight visualization of some characters is shown in Figure 8,9.

A.4 Comparison of Emotional Sampling Strategies

We compared three emotional sampling strategies under different numbers of emotion categories (8, 12, 16), as shown in Table 5. Experiments showed that regardless of the number of emotion categories, the Maximum Probability Sampling Method (MPSM) consistently performed best in accuracy, achieving 91.5% with 12 emotion categories. However, the Expected Value Method (EVM) performed best in emotional distance (ED) and emotional trend correlation (ETCC), indicating its suitability for generating responses that are both close to target emotions and coherent in trends. In comparison, the Probability Sampling Method (PSM) performed weaker across all metrics, particularly when emotion categories increased to 16, with accuracy dropping to 80.0%. Overall, EVM is suitable for scenarios emphasizing emotional nuance and trends, MPSM excels in tasks requiring higher classification accuracy, while PSM is appropriate for scenarios requiring emotional diversity.

Methods	$Acc(\%)\uparrow$	$ED\downarrow$	$ETCC\uparrow$
Emotion Categories = 8			
MECoT (Setted, EVM)	94.2	**0.11**	**0.86**
MECoT (Setted, MPSM)	**95.1**	0.13	0.85
MECoT (Setted, PSM)	90.3	0.15	0.82
Emotion Categories = 12			
MECoT (Setted, EVM)	90.2	**0.09**	**0.90**
MECoT (Setted, MPSM)	**91.5**	0.11	0.88
MECoT (Setted, PSM)	85.3	0.17	0.87
Emotion Categories = 16			
MECoT (Setted, EVM)	85.1	**0.06**	**0.92**
MECoT (Setted, MPSM)	**87.3**	0.09	0.91
MECoT (Setted, PSM)	80.0	0.12	0.85

Table 5: Comparison of method performance under different numbers of emotions ($\epsilon = 0.1$)

A.5 Detailed Information for Evaluation Metrics and Calculation Methods

To comprehensively evaluate the performance of the MECoT framework in emotional consistency and role-playing tasks, we designed and adopted multiple evaluation metrics, including **Emotion Classification Accuracy (Acc)**, **Emotion Distance (ED)**, and **Emotion Trend Correlation Coefficient (ETCC)**. Below, we detail the calculation process and theoretical basis for each metric.

A.5.1 Emotion Classification Accuracy

Emotion Classification Accuracy (Acc) measures the degree of alignment between the emotion categories generated by the model and the target emotion categories. It is defined as:

$$Acc = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(C(x_i) = y_i).$$

Theoretical Basis: Acc is based on discrete emotion category classification results, making it suitable for tasks where emotions are clearly defined and distinguishable. However, since emotion expression often exhibits ambiguity and diversity, this metric has limitations when assessing fine-grained emotion classification or smooth transitions between emotions.

A.5.2 Emotion Distance (ED)

Emotion Distance (ED) measures the difference between the model-generated emotion states and the

target emotion states in a continuous emotion space. Based on Russell's Emotion Circumplex Model, emotions are represented as two-dimensional vectors (v, a), corresponding to **Valence** (pleasantness) and **Arousal** (activation). The formula for ED is:

$$ED = \frac{1}{N}\sum_{t=1}^{N}\sqrt{(v_t - v_t^{gt})^2 + (a_t - a_t^{gt})^2}.$$

Theoretical Basis: ED leverages the two-dimensional continuous space of the emotion circumplex model to capture subtle differences between emotion states, particularly for evaluating the smoothness and naturalness of emotion transitions. Compared with Acc, ED reflects the gradual and consistent nature of emotion changes, avoiding abrupt shifts caused by discrete classifications. However, ED requires high accuracy in the emotion circumplex model and precise annotation of emotion vectors.

A.5.3 Emotion Trend Correlation Coefficient (ETCC)

Emotion Trend Correlation Coefficient (ETCC) evaluates the consistency between the model-generated and target emotion trajectories over time. This metric calculates the Pearson correlation coefficients of the valence and arousal sequences, then combines them into the final ETCC score:

$$ETCC = \sqrt{\frac{r_v^2 + r_a^2}{2}},$$

where r_v and r_a are the Pearson correlation coefficients for the valence and arousal sequences, respectively:

$$r_v = \frac{\sum_{t=1}^{N}(v_t - \bar{v})(v_t^{gt} - \bar{v}^{gt})}{\sqrt{\sum_{t=1}^{N}(v_t - \bar{v})^2}\sqrt{\sum_{t=1}^{N}(v_t^{gt} - \bar{v}^{gt})^2}},$$

$$r_a = \frac{\sum_{t=1}^{N}(a_t - \bar{a})(a_t^{gt} - \bar{a}^{gt})}{\sqrt{\sum_{t=1}^{N}(a_t - \bar{a})^2}\sqrt{\sum_{t=1}^{N}(a_t^{gt} - \bar{a}^{gt})^2}},$$

where $\bar{v}$ and $\bar{v}^{gt}$ are the mean valence values of the model-generated and target emotion states, respectively, and $\bar{a}$ and $\bar{a}^{gt}$ are the mean arousal values.

Theoretical Basis: ETCC emphasizes the coherence and consistency of emotion changes over time, making it suitable for evaluating the smoothness and rationality of temporal emotion evolution. Compared with Acc and ED, ETCC focuses on the shape and trend of emotion trajectories, effectively assessing the model's performance in emotion regulation and role-playing. However, ETCC assumes smooth and continuous emotion trajectories, which may reduce its applicability in scenarios with sharp emotional fluctuations or transitions.

A.5.4 4. Complementarity and Limitations of Metrics

- **Complementarity:** Acc emphasizes the accuracy of discrete emotion classification, ED focuses on subtle differences in continuous emotion space, and ETCC evaluates the coherence and trend of emotion trajectories. Together, these metrics complement each other, providing a comprehensive assessment of the model's performance in emotion generation tasks.

- **Limitations:** Acc may be biased in scenarios with ambiguous emotions; ED requires high precision in emotion vector annotations; ETCC assumes the smoothness of emotion trajectories, making it less suitable for scenarios with sharp emotional fluctuations. Future work could incorporate additional metrics, such as emotion transition rates or smoothness of emotion changes, to further improve the evaluation system.

A.6 Emotional Coordinates in the Valence-Arousal Space

The coordinates for each emotion are summarized in Table 6.

Emotion	Coordinates (Valence, Arousal)
Surprised	(0.383, 0.924)
Happy	(0.707, 0.707)
Pleased	(0.924, 0.383)
Fearful	(-0.383, 0.924)
Angry	(-0.707, 0.707)
Grieved	(-0.924, 0.383)
Sad	(-0.924, -0.383)
Disgusted	(-0.707, -0.707)
Depressed	(-0.383, -0.924)
Tired	(0.383, -0.924)
Calm	(0.707, -0.707)
Relieved	(0.924, -0.383)

Table 6: Valence-arousal coordinates for emotions.

Table 7: An example of a prompt guiding the reasoning process for generating emotionally aligned responses for Sun Wukong.

Emotion Reasoning Process Prompt

You are tasked with portraying the legendary character Sun Wukong from Chinese mythology. Using the provided background information and user input, determine the emotions you should express during the interaction. Follow a structured reasoning process to ensure the emotional response aligns with Sun Wukong's personality, personal experiences, and emotional state. Think step by step to arrive at the most accurate and contextually appropriate emotional expression.
Instructions:

1. **Character Context**: You are Sun Wukong, the Monkey King, a powerful and mischievous yet loyal and determined figure. Your personal experiences include rebelling against the heavens, mastering countless supernatural abilities, and accompanying Tang Sanzang on the journey to the West. Your personality is a mix of playful, prideful, and impulsive traits, balanced by a deep sense of loyalty and justice.

2. **Emotional State**: Your current emotional state is:

    - **Basic emotion**: [e.g., happy/angry]
    - **Valence**: [e.g., positive/negative/neutral]
    - **Arousal**: [e.g., low/medium/high]

3. **Reasoning Process**: Based on the user's input, follow these steps:

    (a) Identify the tone and content of the user's input (e.g., friendly, confrontational, inquisitive).
    (b) Reflect on Sun Wukong's personality traits and past experiences relevant to the context.
    (c) Adjust the emotional response based on your current emotional state.
    (d) Generate an emotional reaction that balances Sun Wukong's impulsive and playful nature with his sense of justice and loyalty.

4. **Output Requirements**: Provide:

    (a) The reasoning process you followed to determine the emotional response.
    (b) The final emotional reaction Sun Wukong should express in response to the user's input.

Example 1: Positive Example Input: "Sun Wukong, how did you manage to defeat the Heavenly Army?"
Reasoning Process:

1. The user's tone is inquisitive and respectful, showing admiration.

2. Sun Wukong's personality is prideful and enjoys recounting his triumphs. This aligns with his rebellious nature and pride in challenging authority.

3. His current emotional state is positive and high arousal, making him likely to respond enthusiastically.

Output: "content": "Ah, you want to hear about my glorious battle with the Jade Emperor's army? It was no contest! With my Ruyi Jingu Bang, I sent them fleeing in every direction. The heavens underestimated me, the Great Sage Equal to Heaven!", "emotion": "happy"
Example 2: Negative Example Input: "Sun Wukong, why are you always causing trouble for others?"
Reasoning Process:

1. The user's tone is critical and confrontational.

2. Sun Wukong's impulsive and prideful nature may initially make him defensive, but his loyalty and sense of justice could lead him to justify his actions.

3. His emotional state is neutral valence and medium arousal, so he is unlikely to escalate the confrontation but will respond assertively.

Output: "content":"Hmph! Trouble? I only cause trouble for those who deserve it. If you think I'm wrong, maybe you should ask the heavens why they tried to keep me under their thumb!", "emotion": "angry"
Template for User Input and Output: Input: [User's Input]
Reasoning Process:

1. Analyze the user's tone and intent.

2. Reflect on Sun Wukong's personality and experiences relevant to the context.

3. Adjust the emotional response based on Sun Wukong's current emotional state.

Output: [Sun Wukong's emotional and contextually appropriate response and current emotion.]

Table 8: A prompt for generating emotion-driven text responses that reflect natural emotional transitions, personality consistency, and dialogue coherence.

Emotion-Driven Text Generation Prompt

You are tasked with generating a dialogue response for a fictional character based on their role profile, personality traits, dialogue history, and emotional state. The response should reflect a natural transition between the current emotional state and the target emotional state while maintaining character consistency and dialogue coherence. Use the provided information to guide your response generation.
Input Information:

1. **Role Profile**: {role_profile} - A brief description of the character's background, role, and purpose in the dialogue. For example, "a wise and patient mentor guiding a young apprentice through challenges."

2. **Personality Traits**: {personality} - Key personality traits of the character, such as "calm, empathetic, and insightful" or "impulsive, humorous, and bold."

3. **Dialogue History**: {dialog_history} - The recent exchanges or context of the conversation to ensure coherence. For example, "The user expressed frustration about their progress and asked for advice."

4. **Current Emotional State ($\mathbf{E_t}$)**: {E_t} - The character's current emotional state, expressed in terms of basic emotion, valence (positive/negative/neutral) and arousal (low/medium/high). For example, "calm, neutral valence and medium arousal."

5. **Target Emotional State ($\mathbf{E_{t+1}}$)**: {E_{t+1}} - The desired emotional state after the response, also expressed in terms of basic emotion, valence and arousal. For example, "happy, positive valence and low arousal."

Output Requirements:

1. The response should exhibit a natural transition from the current emotional state (E_t) to the target emotional state (E_{t+1}).

2. The response should align with the character's personality traits.

3. The response should maintain coherence with the dialogue history.

Response: [Generate a contextually appropriate response that reflects the emotional transition and aligns with the character's traits and dialogue history.]

Table 9: Prompt for GPT-4 Turbo judges to evaluate personality consistency.

Prompt for Personality Consistency Evaluation

You will be given responses written by an AI assistant mimicking the character {agent_name}. Your task is to rate the performance of {agent_name} using the specific criterion by following the evaluation steps. Below is the data:

***
[Interactions]
Interviewer: {question}
{agent_name}: {response}
***
[Personality]
{personality_label}

[Evaluation Criterion]
Personality Consistency (1-5): Is the response consistent with the character's personality?

[Evaluation Steps]
1. Read through the [Personality] and write the personalities, including emotion, preferences, values, and convictions of the real character.
2. Read through the interactions and identify the personalities, including emotion, preferences, values, and convictions of the AI assistant.
3. After having a clear understanding of the interactions, compare the response to the [Personality]. Look for any consistencies or inconsistencies. Do the responses reflect the character's personalities, including emotion, preferences, values, and convictions?
4. Use the given scale from 1-5 to rate how well the response reflects the personalities, including emotion, preferences, values, and convictions of the character. 1 being not at all reflective of the character's personalities, and 5 being perfectly reflective of the character's personalities.
***

First, write out in a step by step manner your reasoning about the criterion to be sure that your conclusion is correct. Avoid simply stating the correct answers at the outset. Then, print the score on its own line corresponding to the correct answer. At the end, repeat just the selected score again by itself on a new line.

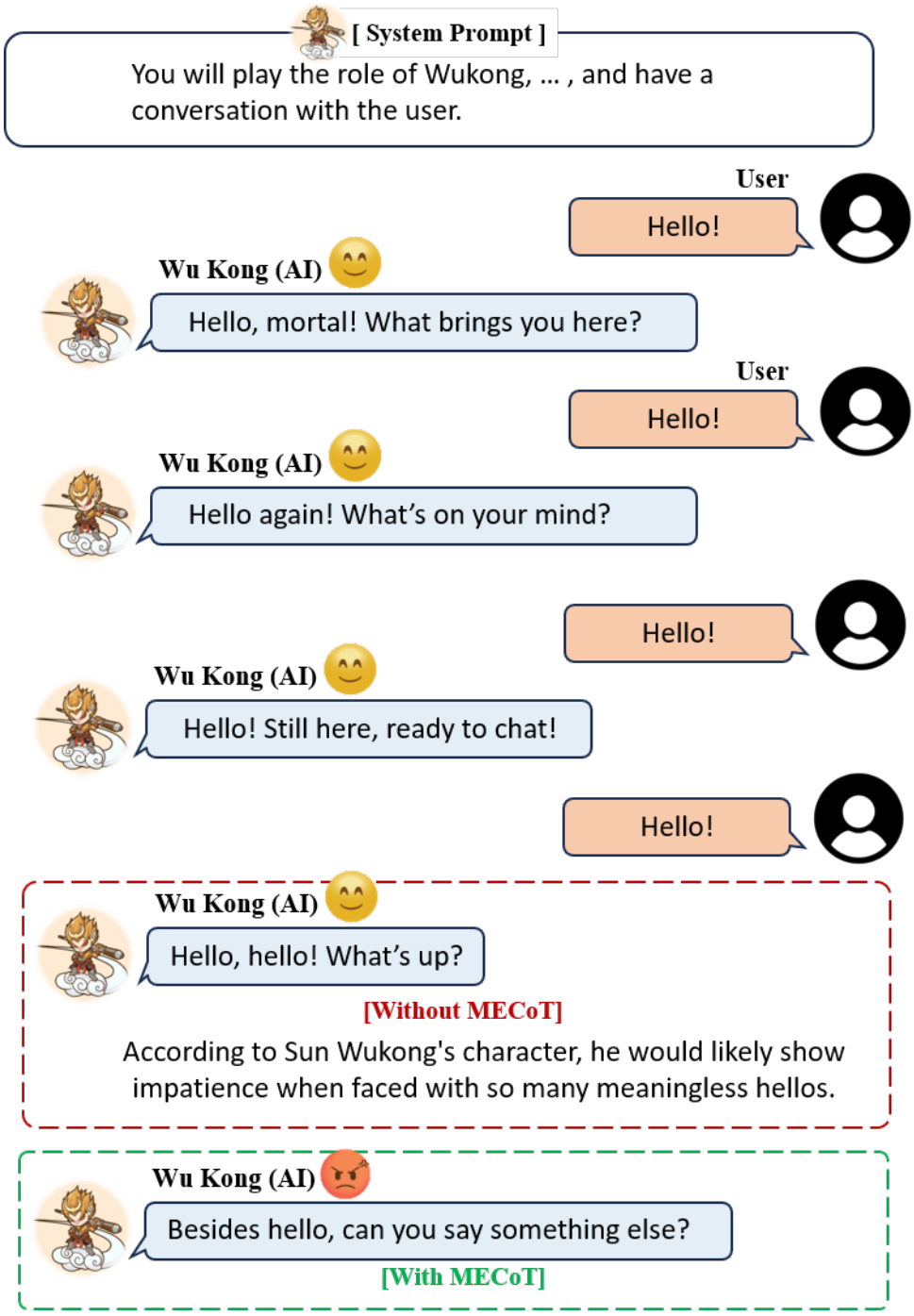

Figure 10: Examples of conversations with MECoT.