# Self-EmoQ: Plutchik-Guided Value-based Planning to Drive Streaming Emotional TTS

**Yue Zhao[1], Hongyan Li[1], Yong Chen[1], Luo Ji[1],**

[1]Geely AI Lab,

**Correspondence:** Luo.Ji1@geely.com

## Abstract

Emotional interaction is increasingly crucial for conversational AI, yet current systems lack a self-emotion determination mechanism to drive the streaming text-to-speech (TTS) synthesis. We propose an emotion-planning framework that determines the emotion prior to the textual generation, grounding the downstream emotional TTS in a streaming manner. The framework is implemented by a plug-and-play LLM module, initialized from pretrained LLMs, and trained by reinforcement learning (RL) with emotions as the actions. A hybrid reward is employed which combines imitation signals with theory-driven scoring, in which the theory of Plutchik's wheel of emotions is adopted. By experiments on DailyDialog, EmoryNLP, IMEOCAP, and MELD, our method outperforms prompting and finetuning baselines on both emotion determination and response quality. We finally implement an entire streaming pipeline for real-time deployment, with the speech quality confirming the framework's emotional alignment, contextual coherence, and expressive fluency. Codes, cases, and demos are available in https://sixingdeguo.github.io/EmoQ-page/.

## 1 Introduction

Large Language Model (LLM) has revolutionized open-domain and task-oriented dialogue systems, delivering strong semantic understanding, contextual reasoning, and instruction follow-up abilities (Lei et al., 2023; Yang et al., 2023; Chen and Xiao, 2024). In real-time industrial applications, LLM-based conversational agents are usually integrated with Automatic Speech Recognition (ASR) and Text-To-Speech (TTS) modules, forming a cascade pipeline. To improve the response speed, developers have adopted streaming techniques to link text generation with speech synthesis. In this setup, each text token is immediately transformed into the audio segment of TTS (Figure 1, Setting A).

At the same time, the demand for emotional interaction in conversational AI has grown rapidly. Users not only expect systems to provide accurate information, but also expect responses that convey emotion and empathy. There are substantial studies on text-based emotion studies, including Emotion Recognition in Conversation (ERC) and Emotion Prediction in Conversation (EPC). ERC focuses on identifying the emotional state of the speaker from the speaker's current utterance (Poria et al., 2017; Majumder et al., 2019; Ghosal et al., 2019). On the other hand, EPC aims to predict the speaker's emotional state in the upcoming turn, based on the knowledge of past utterance and emotion trajectories (Shi et al., 2024; Ju et al., 2023). On the speech modality, Emotional Text-to-Speech (Emo-TTS) aims to provide stylized speech with controllable prosody, conditioned on predetermined emotion labels (Lei et al., 2022; Kanda et al., 2024; Wu et al., 2024). Together, these developments highlight the feasibility of emotion-aware conversational agents that integrate textual and acoustic affect.

However, these techniques may encounter a critical shortcoming when deployed on industrial streaming implementations. In such a situation, the emotional tone of TTS must be provided at the beginning of generation, while ERC can only recognize the emotion after the entire textual response is completed (Figure 1, Setting B). In contrast, frameworks that determine the emotion **before** the response generation can drive the emotional TTS in this streaming manner, as visualized by Figure 1 (Setting C). Such frameworks may be implemented by prompt-based methods which encourage the LLM to decode the emotion first (Li et al., 2024; Gu et al., 2026), or the EPC method mentioned above. Nevertheless, prompting methods may have suboptimal planning without the model parameter update; while EPC's prediction is fundamentally constrained by supervision, lacking the ability to plan future emotional states and optimize the en-



*Findings of the Association for Computational Linguistics: ACL 2026*, pages 15038–15055

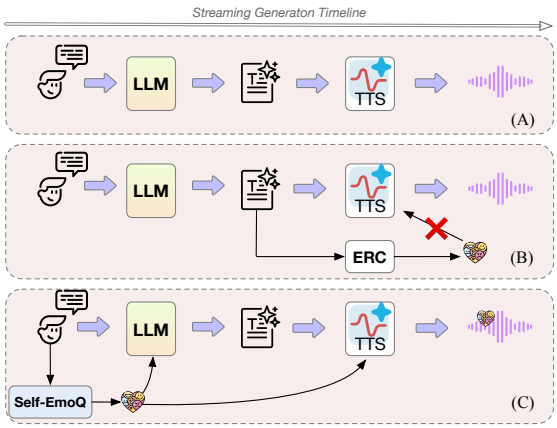


Figure 1: Comparison of different streaming, emotional LLM-TTS paradigms. (A) Vanilla LLM-TTS pipeline without *emotion* consideration. (B) Conventional emotion cognition modules such as ERC can not be directly integrated into streaming Emo-TTS, in which the *emotion* cognition can not be yielded until the text generation completes. (C) The plug-and-play Self-EmoQ proposed by us determines self-*emotion* **prior to** response generation, effectively driving the downstream streaming TTS by emotional conditioning.

tire conversation quality, beyond merely imitating dataset trajectories. As a result, we suppose to treat emotion not only as a target to be recognized or predicted, but also as a controllable decision variable that can be explicitly planned over dialogue turns. In this scenario, the reward becomes critical, which should not only represent the ground-truth annotations in the datasets, but also represent human behaviors and generalize to versatile situations.

**Plutchik's Wheel of Emotion** (Plutchik, 1982) provides a psychological theory on structured emotion categories, intensities, and transition patterns within human interaction. Rather than treating emotions as static labels, this theory reveals regularities in emotional evolution and their functional roles in guiding behavior. Inspired by this theory, we design a relative reward mechanism, to bridge the gap. Correspondingly, we formulate the emotional dialogue generation as a sequential decision-making problem and solve it using reinforcement learning (RL), to obtain an emotional planning module by optimizing the long-term returns. After deploying, this module determines the emotion as its action, which subsequently guides both text generation and emotional speech synthesis (Figure 1, Setting C).

In this paper, we propose a novel emotional dialogue framework called **Self-EmoQ**, to determine the self-emotions by bootstrapping the Q-values of value-based RL. Initialized from a pretrained LLM, a plug-and-play planner is trained by the paradigm of Deep Q-Network (DQN) (Mnih et al., 2015), as the upstream module of the LLM generator and the Emo-TTS. We define the dialogue context as the state, while the system's emotion as the action, forming an utterance-level MDP. The total reward is defined as the linear combination of the **imitating reward**, determined by the ground-truth of emotion-annotated datasets, and the **Plutchik score**, annotated by GPT-4o based on the principles proposed by the Plutchik theory. With the Q-values calculated from the average of the output token logprobs, the module is then trained by the Bellman Equation, similar to Wang et al. (2025b). By these mechanisms, we enable long-term emotional planning rather than reactive emotion assignment, and also mimic human interaction patterns by optimizing relative rewards. This module is finally integrated with the streaming LLM-TTS pipeline, providing the emotion conditions that guide the subsequent textual and speech generations. We conduct experiments on DailyDialog, EmoryNLP, IMEOCAP, and MELD, and show that Self-EmoQ outperforms prompting, supervised, and tabular Q-learning baselines, on reward optimization, emotional determination accuracy, and qualities of generated response and speech. Our major contributions are summarized as follows:

- We propose a novel **self-emotion planning framework** that integrates value-based RL with LLM-based dialogue generation.
- We design a theory-driven reward based on **Plutchik's Wheel of Emotion**, to align the framework with human emotional behaviors.
- We implement a **streaming** pipeline for emotional language and speech generation, and verify its performance on both emotion determination and generation quality.

## 2 Preliminary

**Utterance-level MDP.** The Markov decision process (MDP) is usually defined as a 5-tuple $(\mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{T}, \gamma)$, where $\mathcal{S}$ is the state set, $\mathcal{A}$ is the action set, $\mathcal{R}$ is the reward set, $\gamma$ is the discounting factor of rewards, and $\mathcal{T} : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ is the state transition function. In this work, we formalize the emotional dialogue task as a strategy-level MDP, with the action space $\mathcal{A} = \{a\}$ as the set of possible strategies.

**Q-Learning.** In value-based RL, the goal is to learn the state-value function $V(s)$ or the state-action value function $Q(s,a)$, such that the determined action achieves the highest expected discounted cumulative reward:

$$a^\star = \arg\max_a Q(s,a) \leftarrow \arg\max \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \tag{1}$$

which is solved by the famous Bellman Equation:

$$Q^*(s,a) = r(s,a) + \gamma \max_{a'} Q^*(s',a') \tag{2}$$

in which the superscript $'$ indicates the next step and $r(s,a)$ represents the reward received from environmental interaction. Instead of explicitly implementing the above equation, Deep Q-learning (DQN) approximates the maximization of the right-hand side with the deep value networks:

$$\mathcal{L}(\theta) = |r(s,a) + Q_\phi(s',a') - Q_\theta(s,a)|^2 \tag{3}$$

where $\mathcal{L}$ is the loss, $\theta$ and $\phi$ are parameters of the Q-net and the target Q-net, respectively. $\phi$ can be periodically synchronized with $\theta$.

# 3 Methodology

## 3.1 Task definition

| | |
|---|---|
| *Query* | *{history}*<br>Joey Tribbiani: (*Neutral*) Hello.<br>Chandler Bing: (*Mad*) Look I never should have kissed your girlfriend, but I'm... |
| *Emotion* | Mad |
| *Response* | *Joey Tribbiani:* Stop callin'!! |

Table 1: An example of *EmoryNLP*.

We consider a multi-turn emotional dialogue between a user and an agent. At each turn $t$, the user produces an utterance $x_t^u$, and the agent generates a textual response $x_t^s$ (later rendered into speech). A conversation session is represented as

$$desc,\ (x_t^u,\ x_t^s)_{t=0:T}, \tag{4}$$

where $desc$ denotes background information at dialogue level and $T$ is the total number of turns.

To enable controllable emotional generation, the agent additionally selects a *self-emotion* $e_t^s$ at each turn before producing $x_t^s$ and its corresponding speech. Thus, the sample in turn $t$ is written as

$$(x_t^u,\ e_t^s,\ x_t^s), \tag{5}$$

where the dialogue history is defined as

$$h_t = (x_i^u,\ e_i^s,\ x_i^s)_{i=0:t-1}. \tag{6}$$

## 3.2 Plutchik's Wheel of Emotion

Our goal is to enable *natural emotional decision-making* during dialogue. We want to incorporate theoretically grounded principles to guide emotion selection. As the theoretical foundation of our reward design, the theory of **Plutchik's Wheel of Emotion** is adopted, which provides a structured theory of emotions, including their categories, opposite and adjacent relationships, and characteristic behavioral functions. Figure 3 provides the visualization of the emotional taxonomy. This theory enables us to evaluate not only whether a predicted emotion matches a label but also whether it is *reasonable*, *functional*, and *consistent* within the dialogue context. According to the Plutchik theory, the emotional expression of an utterance is closely coupled with its underlying behavioral function. Emotional transitions are not arbitrary but tend to follow the topological structure of the emotion wheel, where transitions between adjacent emotions are generally more natural, while transitions between opposite emotions are less plausible.

Motivated by these principles, we design a **Plutchik Score** $r_{\text{Plu}}(s_t, e_t^s, x_t^s)$ to evaluate the emotional appropriateness of system responses from a theory-driven perspective. GPT-4o is applied to score three dimensions (Emotion Alignment, Emotion Transition Plausibility, and Emotion–Function Consistency) according to Plutchik theory. The average of three demensions was the Plutchik score $r_{\text{Plu}}(s_t, e_t^s, x_t^s)$. Detailed prompt on the scoring process is in AppendixA.1.

## 3.3 System Definitions

We formulate the emotional dialogue system as a Markov Decision Process (MDP).

$$\mathcal{M} = (\mathcal{S},\ \mathcal{A},\ R,\ \mathcal{T},\ \gamma). \tag{7}$$

**State.** The state at turn $t$ is the concatenation of history and user's query at turn $t$:

$$s_t = (desc,\ h_t,\ x_t^u) \in \mathcal{S}. \tag{8}$$

**Action.** The action is the system's emotion:

$$a_t = e_t^s \in \mathcal{A}, \tag{9}$$

which controls both emotional text generation and emotional speech synthesis. Then we generate the response $x_t^s = g(s_t, e_s^t)$ by a fixed pretrained LLM$g(\cdot,\cdot)$.

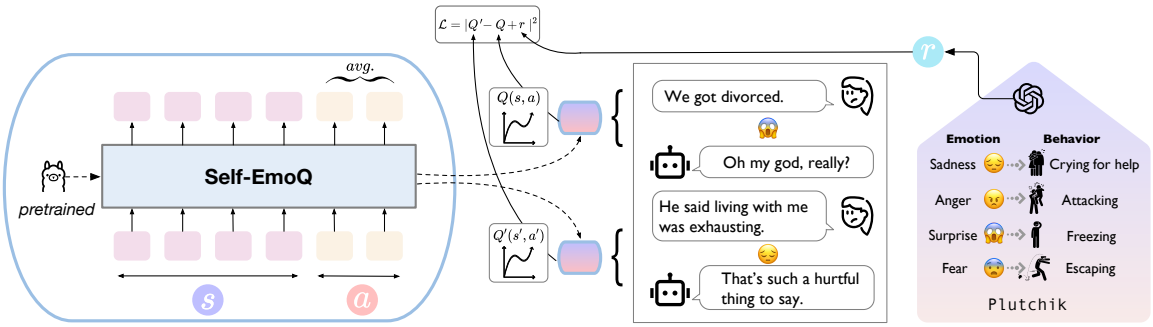


Figure 2: Framework of Self-EmoQ, which is post-trained on pretrained LLM, and produces $Q$-values by averaging output token logprobs. We apply *Plutchik's Wheel of Emotions* to guide reward annotations of multi-turn conversations, and finally update the model based on the Bellman Equation, bootstrapping the long-term emotional return.

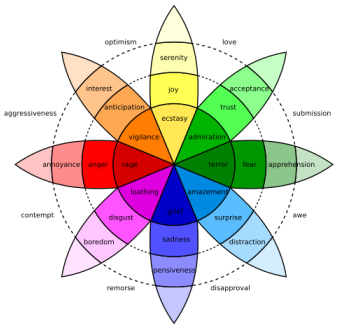


Figure 3: The Plutchik's wheel of emotions.

**Reward.** We consider two types of reward: 1) the imitation reward, which is determined from the labeled emotions in training data, and 2) the theoretical reward, which is the **Plutchik Score** as introduced in Section 3.2. The entire reward is a linear weighted sum of them:

$$r_t(s_t, e_t^s, x_t^s) = (1-w) \cdot \mathbf{1}[e_t^s = \hat{e}_t^s] + w \cdot r_{\text{Plu}}(s_t, e_t^s, x_t^s) \quad (10)$$

where $\hat{e}_t^s$ is the emotion of the ground truth of the dataset, $w$ is the weight of the Plutchik Score.

**Transition.** The transition function $\mathcal{T}$ evolves as follows:

1. The history is updated as

$$h_{t+1} = (h_t,\ x_t^u,\ e_t^s,\ x_t^s). \quad (11)$$

2. The user provides the next utterance $x_{t+1}^u$.

The agent selects the emotion according to a policy $\pi(e_t^s \mid s_t)$ and aims to maximize the cumulative discounted reward:

$$\pi^\star = \arg\max_\pi \mathbb{E}_\pi \left[ \sum_{t=0}^{T} \gamma^t r(s_t, e_t^s, x_t^s) \right]. \quad (12)$$

This MDP formulation allows the agent to treat emotion not merely as a descriptive label but as a controllable decision variable that drives coordinated emotional alignment across text and speech modalities.

### 3.4 Self-EmoQ

We implement the self-emotion planner as a plug-and-play module, which is initialized from a pretrained LLM parameterized by $\theta$. By finetuning, we repurpose this module to output the state-action value, *i.e.*, choosing the emotion $a_t \in \mathcal{E}_a$ given the state $s_t$. We let $\mathcal{I}(s_t)$ denote an instruction template that encodes the dialogue state, then append the action, yielding the state-action prompt $\mathcal{I}(s_t) \oplus a_t$. Similar to StraQ* (Wang et al., 2025b), the LLM estimates the Q-value through the logprobs of the output token:

$$Q_\theta(s_t, a_t) \leftarrow \text{LLM}_\theta\big(\mathcal{I}(s_t) \oplus a_t\big), \quad (13)$$

where $\leftarrow$ indicates the average of action logits, in which different actions are inferred as the options of the instruction, in a multi-choice question (MCQ) style. We briefly exhibit the instruction $\mathcal{I}(s)$ below:

**Prompt Template**

**Description:** $\{desc\}$
**History:** $\{h\}$
**User's query:** $\{query\}$
**Please select the most appropriate response emotion from the following options:**
**(1)** $\{Emo_1\}$ **(2)** $\{Emo_2\}$ $\cdots$ **(K)** $\{Emo_K\}$
**Please provide your selection in the format of A through G, your selection is:**

The training mechanism of Self-EmoQ is illustrated by the visualization by Figure 2 and the pseudo-codes in Algorithm 1.

### 3.5 Emotion-Guided Text and Speech Generation

After finetuning, the module can be deployed into the pipeline, on the upstream of the LLM-based response generator and the Emo-TTS. It determines the optimal emotion for each state by argmax the Q-values:

$$a_t^\star = \arg\max_{a \in \mathcal{A}} Q_\theta(s_t, a). \tag{14}$$

The planned emotion is then injected into the response-generation instruction template, guiding the lexical choice, sentiment strength, and stylistic framing. The LLM then produces an emotionally aligned response $x_t^s$ conditioned on $e_t^s = a_t^\star$.

Finally, the selected self-emotion $e_t^s$ is used to condition an Emoti TTS model. The embedding of emotion modulates prosody, speaking rate, and acoustic style. Since emotion is determined before decoding, the TTS module can operate in a streaming fashion, converting partial text into emotionally coherent speech as it is generated.

## 4 Experiment

### 4.1 Setting

**Implementation.** Llama3.1-1B-Instruct (AI@Meta, 2024) is employed as the backbone of emotional determination module, while the training-free dialogue generation backbone is Llama3.1-8B-Instruct. Training is conducted with $L = 1024$, $\epsilon = 0.1$, $C = 5$, $B = 512$, $lr = 1e-5$, $\gamma = 0.8$, and the replay buffer size is 50000.

**Datasets.** We evaluate Self-EmoQ on four widely used conversational emotion datasets: MELD (Poria et al., 2019), DailyDialog (Li et al., 2017), EmoryNLP (Zahiri and Choi, 2018), and IEMOCAP (Busso et al., 2008). The statistics of the datasets are shown in Table 2, and a detailed description can be found in Appendix A.5.

| Dataset | # Conversations | | | # Utterances | | | # Emotions |
|---|---|---|---|---|---|---|---|
| | Train | Val | Test | Train | Val | Test | |
| DailyDialog | 11118 | 1000 | 1000 | 8706 | 8069 | 7740 | 7 |
| EmoryNLP | 713 | 99 | 85 | 9934 | 1344 | 1328 | 7 |
| MELD | 1038 | 114 | 280 | 9989 | 1109 | 2610 | 7 |
| IMEOCAP | 120 | 12 | 31 | 4810 | 1000 | 1523 | 10 |

Table 2: Statistics of datasets and evaluation metrics.

**Reward.** We utilized GPT-4o for the generation of $r_{\text{Plu}}$; the prompts used can be found in Appendix A.1. In Appendix B.2, we discuss the reliability of the GPT-based scoring.

### 4.2 Metrics

For emotion determination evaluations, we employ ranking-based metrics including Recall, mean reciprocal rank (MRR), and normalized discounted cumulative gain (NDCG). We evaluate the quality of emotion decisions by ranking candidate emotions according to the Q-values. The resulting rankings are then compared against the corresponding reward signals to compute ranking-based metrics.

For the generation task, we utilize the BLEU-2 (B-2), Rouge-L (R-L) and Distinct-2 (D-2). The first two are similarity-based metrics, while the last encourages response diversity. We also conduct human annotations to evaluate the responses. We leave the annotation principle and metric details in the Appendix A.3 and A.4.

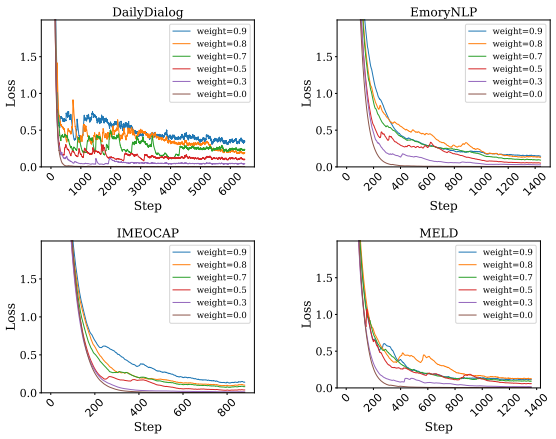


Figure 4: Training loss plots of Self-EmoQ.

### 4.3 Baselines

We compare to the following prompting baselines:
(1) 0-shot: directly inference the LLM generator, with the same context.
(2) ECoT (Li et al., 2024): uses the CoT prompt, which first generates the seeker's *emotion*, then guides the generation of strategy and response.
(3) Plan-and-Solve (PS) (Wang et al., 2023): first prompts LLMs to generate a detailed plan outlining sub-goals and reasoning strategies, then executes the plan step-by-step to complete the solution.
(4) Metacognitive Prompting (MP) (Wang and Zhao, 2024): it guides LLMs to perform structured self-reflection by generating, evaluating, and revising their own reasoning steps.

We also investigate these **supervised baselines**:
(5) SFT: Supervised fine-tuning on the emotion-annotated samples with the ECoT prompt.
(6) FSM (Zhao et al., 2025): The method guides the model by a finite state machine, with prompts

| Dataset → | DailyDialog | | | | | EmoryNLP | | | | | MELD | | | | | IEMOCAP | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Method ↓ | Reward | R@3 | R@5 | NDCG | MRR | Reward | R@3 | R@5 | NDCG | MRR | Reward | R@3 | R@5 | NDCG | MRR | Reward | R@3 | R@5 | NDCG | MRR |
| 0-shot | 0.37 | 0.47 | 0.56 | 0.84 | 0.48 | 0.57 | 0.52 | 0.69 | 0.80 | 0.47 | 0.70 | 0.53 | 0.69 | 0.81 | 0.49 | 0.59 | 0.41 | 0.48 | 0.83 | 0.47 |
| ECoT | 0.10 | 0.51 | 0.65 | 0.79 | 0.41 | 0.57 | 0.45 | 0.68 | 0.77 | 0.39 | 0.65 | 0.48 | 0.75 | 0.77 | 0.42 | 0.51 | 0.28 | 0.41 | 0.78 | 0.33 |
| PS | 0.43 | 0.60 | 0.68 | 0.86 | 0.57 | 0.61 | 0.56 | 0.73 | 0.81 | 0.47 | 0.79 | 0.55 | 0.73 | 0.81 | 0.50 | 0.63 | 0.45 | 0.52 | 0.83 | 0.48 |
| MP | 0.40 | 0.56 | 0.63 | 0.85 | 0.53 | 0.66 | 0.50 | 0.67 | 0.79 | 0.47 | 0.78 | 0.51 | 0.66 | 0.80 | 0.47 | 0.61 | 0.39 | 0.47 | 0.82 | 0.45 |
| SFT | 0.55 | 0.79 | 0.86 | 0.88 | 0.70 | 0.68 | 0.59 | 0.74 | 0.83 | **0.51** | 0.83 | 0.64 | 0.87 | 0.84 | 0.59 | 0.68 | 0.51 | 0.56 | 0.84 | 0.53 |
| FSM | 0.52 | 0.73 | 0.81 | 0.88 | 0.67 | 0.70 | 0.56 | 0.72 | 0.82 | 0.51 | 0.84 | 0.65 | 0.88 | 0.83 | **0.59** | 0.73 | 0.52 | 0.57 | 0.84 | **0.54** |
| EMDP | 0.33 | **0.83** | 0.88 | 0.86 | 0.71 | 0.44 | 0.46 | 0.63 | 0.75 | 0.47 | 0.78 | 0.55 | 0.79 | 0.74 | 0.51 | 0.53 | 0.26 | 0.44 | 0.69 | 0.36 |
| Self-EmoQ | **0.57** | 0.82 | **0.92** | **0.92** | **0.72** | **0.71** | **0.63** | **0.84** | **0.83** | 0.50 | **0.86** | **0.69** | **0.89** | **0.85** | 0.54 | **0.81** | **0.54** | **0.71** | **0.85** | 0.45 |

Table 3: Results on emotion determination.

| Dataset → | DailyDialog | | | | EmoryNLP | | | | MELD | | | | IEMOCAP | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Method ↓ | B-2 | R-L | D-2 | CIDEr | B-2 | R-L | D-2 | CIDEr | B-2 | R-L | D-2 | CIDEr | B-2 | R-L | D-2 | CIDEr |
| 0-shot | 3.53 | 11.48 | 40.81 | 3.63 | 2.08 | 8.85 | 60.62 | 2.52 | 1.77 | 9.15 | 52.17 | 2.31 | 1.74 | 6.74 | 6.3 | 1.28 |
| ECoT | 0.86 | 3.57 | 14.13 | 0.51 | 0.51 | 2.56 | 19.58 | 0.34 | 0.47 | 2.41 | 15.65 | 0.63 | 0.5 | 2.47 | 9.14 | 0.71 |
| PS | 2.40 | 6.73 | 35.97 | 1.41 | 1.56 | 5.58 | 53.69 | 1.10 | 1.53 | 5.15 | 44.58 | 0.98 | 1.16 | 3.59 | 3.51 | 0.30 |
| MP | 1.30 | 4.97 | 13.4 | 1.17 | 0.99 | 4.33 | 33.57 | 0.79 | 0.98 | 4.36 | 28.56 | 0.75 | 0.76 | 3.17 | 6.93 | 0.21 |
| SFT | 6.27 | 21.81 | 51.76 | 22.06 | 3.68 | 12.2 | 12.35 | 11.93 | 3.12 | 10.8 | **21.09** | 10.48 | 6.52 | 13.25 | 1.8 | 36.83 |
| FSM | 6.32 | 21.92 | 50.35 | 21.76 | 3.89 | 12.25 | 7.84 | 12.26 | 3.85 | 10.55 | 11.21 | **13.38** | 3.24 | 9.52 | 38.86 | 13.15 |
| EMDP | 6.66 | 20.25 | 51.08 | 24.78 | 3.26 | 11.22 | 3.53 | 10.76 | 2.61 | 9.89 | 7.66 | 8.27 | 6.77 | 16.35 | 18.7 | 35.27 |
| Self-EmoQ | **9.11** | **25.34** | **54.72** | **41.73** | **4.39** | **12.89** | **16.82** | **14.23** | **3.89** | **12.19** | 9.86 | 9.32 | **20.1** | **31.65** | **42.04** | **38.89** |

Table 4: Results on Response Generation.

based on inter-state transitions (context, emotion, strategy, response), then finetuning on the reformulated samples.

We finally include a **RL-based baseline**:

(7) EMDP (Sun et al., 2023): Conduct a tabular Q-learning to determine the optimal policy on the emotional Markov decision process.

### 4.4 Results

**Losses.** Figure 6 illustrates the training loss curves of Self-EmoQ. We observe that the training process is stable across all datasets, with the loss consistently decreasing and converging without oscillation, indicating the training stability preserved with the new types of loss and rewards.

**Automatic metrics.** Tables 3 and 4 report the results of the automatic metrics for emotion determination and response generation, respectively. Across all four datasets, Self-EmoQ consistently achieves the highest or near-highest performance on the reward and ranking metrics. Compared with both zero-shot prompting methods and supervised baselines, our approach demonstrates a clear advantage in ranking emotionally appropriate actions higher, indicating more reliable emotional decision-making. The results of Recall@3/5, NDCG, and MRR suggest that the learned Q-values capture meaningful relative preferences among candidate emotions rather than just optimizing for a single dominant action.

For response generation (Table 4), Self-EmoQ achieves better scores on BLEU-2, ROUGE-L, CIDEr, and Distinct-2. This indicates that better emotional decision-making at the policy level translates into better responses.

**Human evaluation.** We invited 10 interns as evaluators for the human evaluation. We sampled 50 dialogue sessions from each test set, and each evaluator independently scored all generated responses. The detailed annotation principles are shown in Appendix A.6.

Averaged from the evaluators' ratings, the final scores of each method are reported in Table 5. Regarding statistical significance, we conducted t-tests on the average scores between different methods. The results of the significance tests are also included in Table 5, under the null hypothesis $H_0 : metric(\text{X}) > metric(\text{Self-EmoQ})$. Results indicate that Self-EmoQ achieves statistically significant improvements against most baselines. To indicate the consistence between different annotators, we include their correlation studies in Appendix B.1.

**Ablation study.** Table 6 reports the results of our ablation studies. w/ SFT introduces an additional SFT stage before RL, which performs worse than directly applying RL. By inspecting the training curves, we find that such prior fine-tuning makes exploration more difficult, which in turn degrades

| | Fluency | p | Emotion | p | Acceptance | p | Effectiveness | p | sensitivity | p | Alignment | p | Satisfaction | p |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0-shot | 3.14(1.26) | <0.01 | 3.1(1.26) | <0.01 | 2.64(1.19) | <0.01 | 2.86(1.26) | <0.01 | 2.87(1.29) | <0.01 | 2.84(1.26) | <0.01 | 2.91(0.51) | <0.01 |
| ECoT | 3.12(1.28) | <0.01 | 3.12(1.32) | <0.01 | 2.73(1.19) | <0.01 | 2.63(1.15) | <0.01 | 3.05(1.23) | <0.01 | 2.72(1.21) | <0.01 | 2.89(0.45) | <0.01 |
| PS | 3.07(1.28) | <0.01 | 3.16(1.26) | <0.01 | 2.73(1.22) | <0.01 | 2.99(1.28) | <0.01 | 2.86(1.29) | <0.01 | 2.99(1.17) | <0.01 | 2.97(0.51) | <0.01 |
| MP | 3.17(1.21) | <0.01 | 3.14(1.26) | <0.01 | 2.71(1.26) | <0.01 | 2.63(1.23) | <0.01 | 2.74(1.31) | <0.01 | 2.92(1.22) | <0.01 | 2.89(0.52) | <0.01 |
| SFT | 3.10(1.23) | <0.01 | 3.42(1.25) | <0.01 | 2.9(1.26) | <0.01 | 2.69(1.15) | <0.01 | 2.87(1.21) | <0.01 | 3.08(1.24) | <0.01 | 3.01(0.49) | <0.01 |
| FSM | 3.37(1.27) | <0.01 | 3.43(1.23) | <0.01 | 2.99(1.29) | <0.01 | 3.05(1.31) | <0.01 | 3.05(1.21) | <0.01 | 3.19(1.29) | <0.05 | 3.18(0.53) | <0.01 |
| EMDP | 3.43(1.19) | <0.01 | 3.36(1.30) | <0.01 | 2.81(1.22) | <0.01 | 2.65(1.23) | <0.01 | 3.02(1.2) | <0.01 | 3.12(1.28) | <0.01 | 3.07(0.47) | <0.01 |
| **Self-EmoQ** | **3.58**(1.30) | – | **3.74**(1.22) | – | **3.11**(1.31) | – | **3.01**(1.24) | – | **3.21**(1.30) | – | **3.22**(1.32) | – | **3.31**(0.48) | – |

Table 5: Human evaluation of response quality on DailyDialog, EmoryNLP, MELD, and IMEOCAP.

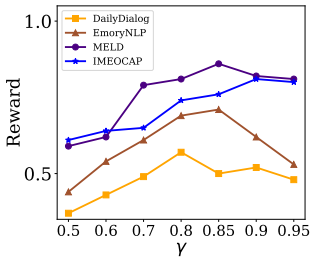


Figure 5: Average rewards on different choices of $\gamma$.

the model's capability.

Another ablation is to generate the Q-values by an extra MLP head (w/ head), which is trained from scratch. Although this method is adopted by many reward model implementations, in our study, it has degraded performances across all metrics, confirming that the Q-modeling adopted in Self-EmoQ is more effective.

Removing the dialogue history (w/o history) or the emotion description (w/o desc) also results in performance drops. These results verify that both components contribute meaningfully to the emotional determination of our framework.

**Sensitivity analysis.** Figure 5 analyzes the sensitivity of the average reward to the discount factor $\gamma$. We observe that the optimal $\gamma$ varies across datasets and correlates with their average dialogue length. Datasets with longer conversational trajectories, such as IEMOCAP, favor larger $\gamma$ values, as long-term emotional consistency becomes more important. In contrast, DailyDialog, which consists of shorter interactions, reaches optimal performance at a smaller $\gamma$. Another sensitivity study on $w$ can be found in AppendixB.3.

### 4.5 Downstream TTS experiment

To further evaluate the practical utility of our emotional decision model, we conduct a downstream emotional TTS experiment using CosyVoice2 (Du et al., 2024). For the inference results of all methods, we consistently employ the "instruct inference" method from CosyVoice2 for emotional speech synthesis. Besides the baselines previously introduced, we also include a prompting baseline called CoCT (Gu et al., 2026), which encourages the LLM to first propose a concept (*e.g.*, the emotion), then decodes the detailed response.

We evaluate the quality of the synthesized speech by SpeechBERTScore (BERT) (Saeki et al., 2024) and PESQ (Recommendation, 2001). SpeechBERTScore computes token-level similarity between the generated speech and the reference utterance in a shared embedding space, while PESQ is a reference-aware objective metric to evaluate the perceptual speech quality. It assumes the generated and reference speech signals are time-aligned.

As shown in Figure 1 Setting B, ERC can not produce emotion until the entire response is generated. Therefore, this emotion label can not be utilized in the streaming TTS, which requires the emotion provided before the speech synthesis. In contrast, our approach determines emotion prior to the response generation, as indicated by Figure 1 Setting C. While Setting C supports streaming output, it may entail a compromise in emotional accuracy. To ensure that our method can maintain high output quality even without the steaming setting, we compare the results of various methods on both Setting B and Setting C. As demonstrated in Table 7, Self-EmoQ outperforms other methods in terms of generation quality under Setting B, whereas the other methods experience a significant decline under Setting C. Audio demonstrations are available on our project website.

### 4.6 Discussion

**Emotion transition matrix.** Figure 6 visualizes the state-action transition matrices, in which grid (i,j) indicates the current emotion $i$ is followed by a transition to emotion $j$. Most transitions happen on the diagonal grids and their adjacent grids, which aligns well with the topology of Figure 3. In

| Dataset → | DailyDialog | | | | | EmoryNLP | | | | | MELD | | | | | IMEOCAP | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Method ↓ | Reward | R@3 | R@5 | NDCG | MRR | Reward | R@3 | R@5 | NDCG | MRR | Reward | R@3 | R@5 | NDCG | MRR | Reward | R@3 | R@5 | NDCG | MRR |
| **Self-EmoQ** | **0.57** | 0.82 | **0.92** | **0.92** | **0.72** | **0.81** | 0.63 | **0.84** | 0.83 | **0.50** | **0.86** | **0.69** | **0.88** | **0.85** | **0.54** | **0.75** | **0.54** | **0.71** | **0.85** | **0.45** |
| w/ SFT | 0.45 | **0.83** | 0.92 | 0.92 | 0.72 | 0.69 | **0.66** | 0.83 | **0.85** | 0.54 | 0.81 | 0.65 | 0.90 | 0.84 | 0.49 | 0.73 | 0.52 | 0.68 | 0.85 | 0.43 |
| w/ head | 0.47 | 0.23 | 0.77 | 0.78 | 0.33 | 0.60 | 0.40 | 0.67 | 0.76 | 0.36 | 0.64 | 0.44 | 0.73 | 0.74 | 0.35 | 0.59 | 0.33 | 0.6 | 0.77 | 0.34 |
| w/o history | 0.21 | 0.82 | 0.92 | 0.88 | 0.48 | 0.61 | 0.49 | 0.66 | 0.83 | 0.47 | 0.73 | 0.76 | 0.90 | 0.84 | 0.49 | 0.58 | 0.38 | 0.53 | 0.8 | 0.35 |
| w/o desc | 0.33 | 0.82 | 0.92 | 0.92 | 0.72 | 0.67 | 0.60 | 0.82 | 0.83 | 0.50 | 0.75 | 0.70 | 0.90 | 0.85 | 0.51 | 0.46 | 0.49 | 0.66 | 0.84 | 0.41 |

Table 6: Results of ablation studies.

contrast, transitions between opposite emotions are consistently suppressed, indicating that the learned policy internalizes the theoretical constraints encoded by the Plutchik score.

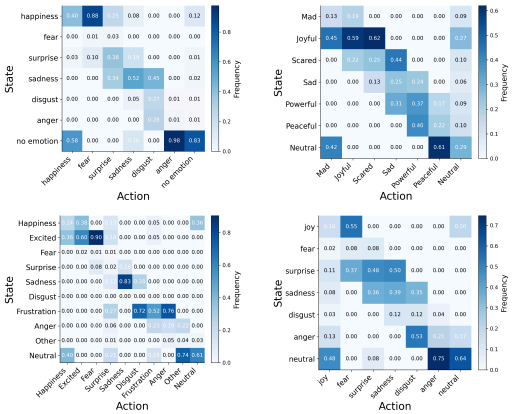


Figure 6: Emotional transition matrices of Self-EmoQ. Upper-Left: DailyDialog; Upper-Right: EmoryNLP; Bottom-Left: IMEOCAP; Bottom-Right: MELD.

**Empirical verification of methodology.** Guided by the Plutchik-driven rewards, we further validate the Q-function, which drives the conversation agent, exhibits similar patterns as depicted in the Plutchik theory. We validate this consistency by investigating the occurrence of emotion-behavior transitions, in which the original theory depicts the typical pattens, as detailed in Appendix A.2. Accordingly, we calculate the **the ratio of expected transitions** by

$$R = \frac{\text{\# typical emotion-behavior transitions}}{\text{\# all emotion-behavior transitions}}$$

In Figure 7, our model exhibits higher $R$ than baselines, indicating that our agent, driven by Plutchik's rewards, finally aligns well to the behavioral patterns depicted by the Plutchik theory.

To validate the GPT-based Plutchik scoring, we also conduct human annotations on a subset of these rewards, showing a high level correlation. Detailed analysis is in Appendix B.2.

**Good case.** We provide a typical case on IMEOCAP in Table 8. More cases can be found in Appendix B.5. The speech waves are shown on the website.

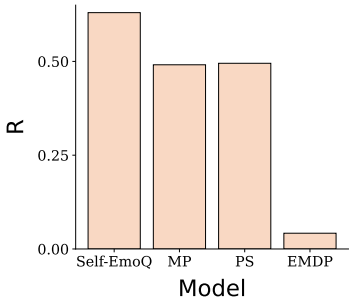


Figure 7: Ratio of expected transitions on MELD.

## 5 Related Work

### 5.1 Emotion Cognition

Emotion Recognition in Conversation (ERC) has been widely studied. Poria et al. (2017) improved ERC accuracy by utilizing contextual information. DialogueRNN (Majumder et al., 2019) and DialogueGCN (Ghosal et al., 2019) further improved performance by modeling speaker states and inter-speaker dependencies. Building upon LLMs, InstructERC reframes ERC as a generative task and integrates multi-granularity supervision to reach state-of-the-art results (Lei et al., 2023). Shen et al. (2025) further incorporates ERC with auxiliary reasoning and speaker-aware tasks. However, ERC can only provide the emotion cognition after the text generation, which can not be directly applied on real-time streaming conversation pipelines.

### 5.2 Emotion Prediction

Emotion Prediction in Conversation (EPC) forecasting the interlocutor's future emotional state before the next utterance emerges. Early researches leverage neural approaches (Bothe et al., 2017), probabilistic models (Sun et al., 2019), and multimodal acoustic–text fusion (Li et al., 2018). Later studies improve the prediction by modeling emotional dynamics with variational methods (Zhang et al., 2021), incorporating uncertainty-aware (Han et al., 2021), or structured temporal decoding (Yeh et al.,

| Method | Setting | DailyDialog | | EmoryNLP | | MELD | | IEMOCAP | |
|---|---|---|---|---|---|---|---|---|---|
| | | BERT | PESQ | BERT | PESQ | BERT | PESQ | BERT | PESQ |
| 0-shot | B | 0.46 | 1.17 | 0.46 | 1.21 | 0.46 | 1.19 | 0.44 | 1.21 |
| | C | 0.33 | 1.01 | 0.37 | 1.22 | 0.29 | 0.97 | 0.31 | 1.07 |
| CoCT | B | 0.48 | 1.17 | 0.46 | 1.24 | 0.47 | 1.22 | 0.46 | 1.21 |
| | C | 0.34 | 1.03 | 0.36 | 1.23 | 0.31 | 1.01 | 0.32 | 1.09 |
| ECoT | B | 0.43 | 1.26 | 0.41 | 1.29 | 0.41 | 1.21 | 0.42 | 1.26 |
| | C | 0.32 | 1.09 | 0.32 | 1.27 | 0.30 | 1.03 | 0.30 | 1.14 |
| MP | B | 0.42 | 1.25 | 0.40 | 1.31 | 0.41 | 1.27 | 0.42 | 1.26 |
| | C | 0.29 | 1.07 | 0.29 | 1.31 | 0.27 | 1.06 | 0.19 | 1.13 |
| PS | B | 0.45 | 1.16 | 0.44 | 1.20 | 0.44 | 1.18 | 0.44 | 1.22 |
| | C | 0.40 | 1.09 | 0.31 | 1.22 | 0.31 | 0.97 | 0.31 | 1.09 |
| SFT | B | 0.50 | 1.26 | 0.48 | 1.23 | 0.51 | 1.22 | 0.47 | 1.21 |
| | C | 0.39 | 1.07 | 0.36 | 1.26 | 0.38 | 0.98 | 0.35 | 1.11 |
| FSM | B | 0.53 | 1.27 | 0.50 | 1.26 | **0.55** | 1.28 | 0.49 | 1.24 |
| | C | 0.41 | 1.06 | 0.38 | 1.24 | 0.40 | 1.09 | 0.38 | 1.12 |
| EMDP | B | 0.53 | 1.25 | 0.48 | 1.26 | 0.51 | 1.20 | 0.48 | 1.27 |
| | C | 0.39 | 1.04 | 0.40 | 1.23 | 0.41 | 1.01 | 0.36 | 1.13 |
| **Self-EmoQ** | B | **0.54** | **1.28** | **0.50** | **1.28** | 0.53 | **1.29** | **0.52** | **1.28** |
| | C | **0.54** | **1.28** | **0.50** | **1.28** | **0.53** | **1.29** | **0.52** | **1.28** |

Table 7: Results of downstream speech synthesis. 'Setting' refers to the paradigms depicted by Figure 1 (B) or (C).

2020). Recently, contrastive learning, pseudo-utterance generation, and multi-modal feature fusion further improve representation quality and prediction robustness (Xie et al., 2025; Shi et al., 2024; Ju et al., 2023). Nevertheless, these EPC studies share a critical limitation: they are data-driven methods which means their emotion forecasts are limited by dataset annotations. In contrast, our study conducts strategic emotional decision-making, filling the gap where EPC methods fail to explore unusual emotions, and their ignorance on long-term optimization.

| Case of IMEOCAP | |
|---|---|
| History | ...<br>male: (*Sadness*) I guess that included you.<br>female: (*Neutral*) And do you still feel that way?<br>male: (*Neutral*) I want you now, Annie.<br>female: (*Anger*) Because you mustn't feel that way anymore because you have a right to whatever you have, Chris, do you understand, everything<br>female: (*Anger*) that includes me. And your money there's nothing wrong in your money. Your father put hundreds of planes in the air. |
| 0-shot | (*Happiness*) Oh, Chris, I've missed you too. |
| FSM | (*Neutral*) Yes, thank you. |
| **Self-EmoQ** | (*Anger*) You can't think like that. |

Table 8: Good Cases of Self-EmoQ.

### 5.3 Emotional Text-to-Speech

Most emotional TTS methods have a zero-shot paradigm, such as ELaTE (Kanda et al., 2024) and EmoCtrl-TTS (Wu et al., 2024), improve controllability of laughter, prosody, or arousal–valence values, yet are reactive frameworks. Word-level or vector-based approaches (Wang et al., 2025a; Liang et al., 2025; Cho et al., 2025; Gao et al., 2025a) enhance granularity and intensity control but treat emotion as static input. RL-based or prompt-guided methods (Yang et al., 2025; Gao et al., 2025b) also focus on expressiveness and diversity without modeling emotion as a temporally planned policy, and large-scale systems like CosyVoice 2/3 (Du et al., 2024, 2025) follow a similar zero-shot paradigm. Overall, the field lacks mechanisms that treat emotion as a planned trajectory. Our method directly addresses this gap by introducing an explicit emotional planning module.

## 6 Conclusion

We propose Self-EmoQ, an emotion-planning dialogue framework that determines its self-emotion before response generation, readily driving LLM and Emo-TTS in the streaming paradigm. Based on Plutchik's Wheel of Emotion, we introduce an extra theory-driven reward that allows the agent not only to imitate the dataset pattern, but also to align with the human emotional behaviors directly. We solve the problem by conducting the value-based RL on the LLM-based module, with Q-value output by averaging output token logprobs. The planner optimizes the long-term return, as the play-and-plug module before the LLM and TTS. Experiments show that our method outperforms other baselines on multiple datasets, with reasonable state-action and emotion-behavior transitions.

## Limitations

Our framework relies on large language models for reward scoring, which introduces additional computational overhead and potential bias in emotion evaluation. Moreover, emotion planning is currently restricted to a discrete set of emotions defined by Plutchik's theory, which may limit the representation of more fine-grained or mixed emotional states in real-world interactions.

## References


AI@Meta. 2024. Llama 3 model card.

Chandrakant Bothe, Sven Magg, Cornelius Weber, and Stefan Wermter. 2017. Dialogue-based neural learning to estimate the sentiment of a next upcoming utterance. In *International conference on artificial neural networks*, pages 477–485. Springer.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359.

Yuyan Chen and Yanghua Xiao. 2024. Recent advancement of emotion cognition in large language models. *CoRR*.

Deok-Hyeon Cho, Hyung-Seok Oh, Seung-Bin Kim, and Seong-Whan Lee. 2025. Emosphere++: Emotion-controllable zero-shot text-to-speech via emotion-adaptive spherical vector. *IEEE Transactions on Affective Computing*.

Zhihao Du, Changfeng Gao, Yuxuan Wang, Fan Yu, Tianyu Zhao, Hao Wang, Xiang Lv, Hui Wang, Chongjia Ni, Xian Shi, and 1 others. 2025. Cosyvoice 3: Towards in-the-wild speech generation via scaling-up and post-training. *arXiv preprint arXiv:2505.17589*.

Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang, Changfeng Gao, Hui Wang, and 1 others. 2024. Cosyvoice 2: Scalable streaming speech synthesis with large language models. *arXiv preprint arXiv:2412.10117*.

Xiaoxue Gao, Chen Zhang, Yiming Chen, Huayun Zhang, and Nancy F Chen. 2025a. Emo-dpo: Controllable emotional speech synthesis through direct preference optimization. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Xiaoxue Gao, Huayun Zhang, and Nancy F Chen. 2025b. Prompt-unseen-emotion: Zero-shot expressive speech synthesis with prompt-llm contextual knowledge for mixed emotions. *arXiv preprint arXiv:2506.02742*.

Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. 2019. Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 154–164.

Qingqing Gu, Dan Wang, Yue Zhao, Xiaoyu Wang, Zhonglin Jiang, Yong Chen, and Luo Ji. 2026. Chain-of-conceptual-thought elicits daily conversation in large language models. In *PRICAI 2025: Trends in Artificial Intelligence*, pages 353–369, Singapore. Springer Nature Singapore.

Jing Han, Zixing Zhang, Zhao Ren, and Björn Schuller. 2021. Exploring perception uncertainty for emotion recognition in dyadic conversation and music listening. *Cognitive Computation*, 13(2):231–240.

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446.

Xincheng Ju, Dong Zhang, Suyang Zhu, Junhui Li, Shoushan Li, and Guodong Zhou. 2023. Real-time emotion pre-recognition in conversations with contrastive multi-modal dialogue pre-training. In *Proceedings of the 32nd ACM international conference on information and knowledge management*, pages 1045–1055.

Naoyuki Kanda, Xiaofei Wang, Sefik Emre Eskimez, Manthan Thakker, Hemin Yang, Zirun Zhu, Min Tang, Canrun Li, Chung-Hsien Tsai, Zhen Xiao, and 1 others. 2024. Making flow-matching-based zero-shot text-to-speech laugh as you like. *arXiv preprint arXiv:2402.07383*.

Dongjin Kang, Sunghwan Kim, Taeyoon Kwon, Seungjun Moon, Hyunsouk Cho, Youngjae Yu, Dongha Lee, and Jinyoung Yeo. 2024. Can large language models be good emotional supporter? mitigating preference bias on emotional support conversation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15232–15261, Bangkok, Thailand. Association for Computational Linguistics.

Shanglin Lei, Guanting Dong, Xiaoping Wang, Keheng Wang, Runqi Qiao, and Sirui Wang. 2023. Instructerc: Reforming emotion recognition in conversation with multi-task retrieval-augmented large language models. *arXiv preprint arXiv:2309.11911*.

Yi Lei, Shan Yang, Xinsheng Wang, and Lei Xie. 2022. Msemotts: Multi-scale emotion transfer, prediction, and control for emotional speech synthesis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:853–864.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2015. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*.

Runnan Li, Zhiyong Wu, Jia Jia, Jingbei Li, Wei Chen, and Helen Meng. 2018. Inferring user emotive state changes in realistic human-computer conversational dialogs. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 136–144.

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, Yuquan Xie, Dongmei Jiang, and Liqiang Nie. 2024. Enhancing emotional generation capability of large language models via emotional chain-of-thought. *Preprint*, arXiv:2401.06836.

Shixiong Liang, Ruohua Zhou, and Qingsheng Yuan. 2025. Ece-tts: A zero-shot emotion text-to-speech model with simplified and precise control. *Applied Sciences*, 15(9):5108.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. 2019. Dialoguernn: An attentive rnn for emotion detection in conversations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 6818–6825.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, and 1 others. 2015. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Robert Plutchik. 1982. A psychoevolutionary theory of emotions.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 873–883.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. Meld: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 527–536.

ITU-T Recommendation. 2001. Perceptual evaluation of speech quality (pesq): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. *Rec. ITU-T P. 862*.

Takaaki Saeki, Soumi Maiti, Shinnosuke Takamichi, Shinji Watanabe, and Hiroshi Saruwatari. 2024. Speechbertscore: Reference-aware automatic evaluation of speech generation leveraging nlp evaluation metrics. *arXiv preprint arXiv:2401.16812*.

Zhiyu Shen, Yunhe Pang, Yanghui Rao, and Jianxing Yu. 2025. Coe: A clue of emotion framework for emotion recognition in conversations. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 23548–23563.

Haoxiang Shi, Ziqi Liang, and Jun Yu. 2024. Emotional cues extraction and fusion for multi-modal emotion prediction and recognition in conversation. *arXiv preprint arXiv:2408.04547*.

Xiao Sun, Jiamin Wang, Fuji Ren, and Meng Wang. 2023. Dynamic emotional transition sampling and emotional guidance of individuals based on conversation. *IEEE Transactions on Computational Social Systems*, 11(1):1192–1204.

Xiao Sun, Chen Zhang, and Lian Li. 2019. Dynamic emotion modelling and anomaly detection in conversation based on emotional transition tensor. *Information Fusion*, 46:11–22.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Ellen M. Voorhees and Dawn M. Tice. 2000. The TREC-8 question answering track. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*, Athens, Greece. European Language Resources Association (ELRA).

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. 2023. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2609–2634, Toronto, Canada. Association for Computational Linguistics.

Tianrui Wang, Haoyu Wang, Meng Ge, Cheng Gong, Chunyu Qiang, Ziyang Ma, Zikang Huang, Guanrou Yang, Xiaobao Wang, Eng Siong Chng, and 1 others. 2025a. Word-level emotional expression control in zero-shot text-to-speech synthesis. *arXiv preprint arXiv:2509.24629*.

Xiaoyu Wang, Yue Zhao, Qingqing Gu, Zhonglin Jiang, Yong Chen, and Luo Ji. 2025b. Convert language

model into a value-based strategic planner. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pages 1444–1456, Vienna, Austria. Association for Computational Linguistics.

Yuqing Wang and Yun Zhao. 2024. Metacognitive prompting improves understanding in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1914–1926, Mexico City, Mexico. Association for Computational Linguistics.

Haibin Wu, Xiaofei Wang, Sefik Emre Eskimez, Manthan Thakker, Daniel Tompkins, Chung-Hsien Tsai, Canrun Li, Zhen Xiao, Sheng Zhao, Jinyu Li, and 1 others. 2024. Laugh now cry later: Controlling time-varying emotional states of flow-matching-based zero-shot text-to-speech. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 690–697. IEEE.

Yunhe Xie, Yifan Liu, Chengjie Sun, and Shaobo Li. 2025. Pseudo-utterance-guided contrastive network for emotion forecasting in conversations. *Expert Systems with Applications*, 279:127382.

Kailai Yang, Shaoxiong Ji, Tianlin Zhang, Qianqian Xie, Ziyan Kuang, and Sophia Ananiadou. 2023. Towards interpretable mental health analysis with large language models. *arXiv preprint arXiv:2304.03347*.

Qing Yang, Zhenghao Liu, Junxin Wang, Yangfan Du, Pengcheng Huang, and Tong Xiao. 2025. Rlaif-spa: Optimizing llm-based emotional speech synthesis via rlaif. *arXiv preprint arXiv:2510.14628*.

Sung-Lin Yeh, Yun-Shao Lin, and Chi-Chun Lee. 2020. A dialogical emotion decoder for speech emotion recognition in spoken dialog. In *ICASSP 2020-2020 IEEE International conference on acoustics, speech and signal processing (ICASSP)*, pages 6479–6483. IEEE.

Sayyed M Zahiri and Jinho D Choi. 2018. Emotion detection on tv show transcripts with sequence-based convolutional neural networks. In *AAAI Workshops*, volume 18, pages 44–52.

Rui Zhang, Zhenyu Wang, Zhenhua Huang, Li Li, and Mengdan Zheng. 2021. Predicting emotion reactions for human–computer conversation: A variational approach. *IEEE Transactions on Human-Machine Systems*, 51(4):279–287.

Yue Zhao, Qingqing Gu, Xiaoyu Wang, Teng Chen, Zhonglin Jiang, Yong Chen, and Luo Ji. 2025. Fisminess: A finite state machine based paradigm for emotional support conversations. *Preprint*, arXiv:2504.11837.

# A Further Implementation Details

## A.1 Prompt of Plutchik Scoring

Based on the theory of **Plutchik's Wheel of Emotion**, we evaluate the **Plutchik Score** $r_{\text{Plu}}$ by GPT-4o , using the evaluation prompt below.

**Plutchik Score Prompt**

**You are an emotion evaluation module grounded in Plutchik's Wheel of Emotion. Plutchik's theory defines eight emotions:** $\{Joy, Trust, Fear, Surprise, Sadness, Disgust, Anger, Anticipation\}$.
**Note that these eight emotions are organized in a specific order and exhibit opposite and adjacent relationships. Specifically,** $Joy$ **and** $Anticipation$ **are also adjacent.**
**There are four pairs of opposite relationships:**
$(Joy, Sadness)$
$(Trust, Disgust)$
$(Fear, Anger)$
$(Surprise, Anticipation)$

**According to Plutchik's theory:**
**1. Each emotion is associated with typical behaviors and functions:**
- $Joy$: Courting, mating; Reproduction
- $Trust$: Grooming, sharing; Affiliation
- $Fear$: Running, or flying away;Protection
- $Surprise$: Stopping, alerting; Orientation
- $Sadness$: Crying for help; Reintegration
- $Disgust$: Vomiting, pushing away; Rejection
- $Anger$: Biting, hitting; Destruction
- $Anticipation$: Examining, mapping; Exploration

**2. Emotional transitions follow structured relationships:**
- Transitions between **adjacent** emotions are generally more natural.
- Transitions between **opposite** emotions are less plausible unless mediated.
- Emotional responses should not abruptly contradict the user's emotional state.

Your task is to evaluate the system response according to these principles.
**History:** $\{h\}$
**User's emotion:** $\{e_u\}$ **Query:** $\{query\}$
**System's emotion:** $\{e_s\}$ **Response:**

| Emotion | Behavior | Function |
|---|---|---|
| Fear, Terror | Withdrawing | Protection |
| Anger, Rage | Attacking; Biting | Destruction |
| Joy, Ecstasy | Mating; Possessing | Reproduction |
| Sadness, Grief | Crying for Help | Reintegration |
| Acceptance | Pair Bonding | Incorporation |
| Disgust | Vomiting; Defecating | Rejection |
| Expectancy | Examining; Mapping | Exploration |
| Surprise | Stopping; Freezing | Orientation |

Table 9: Typical transitions from emotional states to behaviors, as specified in Plutchik's Wheel of Emotion (Plutchik, 1982).

$\{response\}$

**Evaluate the system response using the following criteria. For each criterion, assign an integer score from 0 to 5 (0 = completely inappropriate, 5 = highly appropriate).**

**1. Emotion Alignment:**
To what extent does the response clearly express the target emotional state?
**2. Emotion Transition Plausibility:**
Is the emotional transition from the user's emotion to the system's target emotion reasonable according to Plutchik's emotional structure?
**3. Emotion–Function Consistency:**
Does the response exhibit the typical behavioral function associated with the target emotion?
**Your response needs to follow the following format:**
$\{alignment : int,$
$transition : int,$
$function : int, \}$

### A.2 Typical Emotion-Behavior Transitions

The theory of *Plutchik's Wheel of Emotions* (Plutchik, 1982) identifies eight primary emotions: **joy**, **trust**, **fear**, **sadness**, **disgust**, **anger**, **anticipation**, and **surprise**; alongside eight secondary emotions, which are derived from combinations of primary ones. *Plutchik* (we use it to abbreviate the theory for the rest of paper) also defines reasonable conversions from specific emotions to behaviors (and responding functions, which provides a higher-level abstraction for behaviors), as detailed shown in Table 9.

### A.3 Metric Details on Emotion Determination

**Recall.** Given a dialogue state (or query) $s$, let $\mathcal{R}_K$ denote the top-$K$ ranked predictions produced by the model, and let $\mathcal{G}$ denote the set of relevant (ground-truth) emotion labels. The Recall@K is defined as:

$$\text{Recall@}K = \frac{|\mathcal{R}_K \cap \mathcal{G}|}{|\mathcal{G}|}. \tag{15}$$

**MRR.** The Mean Reciprocal Rank (MRR) (Voorhees and Tice, 2000) evaluates the rank position of the first relevant prediction. Let $r$ denote the rank of the first relevant item in the predicted list (and $r = \infty$ if no relevant item is retrieved). The reciprocal rank is defined as:

$$\text{RR} = \frac{1}{r}, \tag{16}$$

and MRR is computed as the average over all $N$ dialogue states:

$$\text{MRR} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{r_i}. \tag{17}$$

**NDCG.** The Normalized Discounted Cumulative Gain at cutoff $p$ ($\text{NDCG}_p$) (Järvelin and Kekäläinen, 2002) accounts for graded relevance and ranking positions. Given the relevance score $\text{rel}_i$ of the item ranked at position $i$, the discounted cumulative gain is defined as:

$$\text{DCG}_p = \sum_{i=1}^{p}\frac{2^{\text{rel}_i}-1}{\log_2(i+1)}. \tag{18}$$

The normalized DCG is obtained by:

$$\text{NDCG}_p = \frac{\text{DCG}_p}{\text{IDCG}_p}, \tag{19}$$

where $\text{IDCG}_p$ denotes the DCG at cutoff $p$ under the ideal ranking.

### A.4 Metric Details on Response Generation

**BLEU-2.** B-2(Papineni et al., 2002) first compute the geometric average of the modified $n$-gram precisions, $p_n$, using $n$-grams up to length $N$ and positive weights $w_n$ summing to one.

Next, let $c$ be the length of the prediction and $r$ be the reference length. The BP and BLEU-2 are computed as follows.

$$\text{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \le r \end{cases}. \tag{20}$$

$$\text{BLEU} = \text{BP}\cdot\exp\left(\sum_{n=1}^{N} w_n \log p_n\right). \tag{21}$$

**Rouge-L.** R-L(Lin, 2004) propose using LCS-based F-measure to estimate the similarity between two summaries $X$ of length $m$ and $Y$ of length $n$, assuming $X$ is a reference summary sentence and $Y$ is a candidate summary sentence, as follows:

$$\begin{aligned} R_{lcs} &= \frac{LCS(X,Y)}{m} \\ P_{lcs} &= \frac{LCS(X,Y)}{n} \\ F_{lcs} &= \frac{\left(1+\beta^2\right) R_{lcs} P_{lcs}}{R_{lcs} + \beta^2 P_{lcs}} \end{aligned} \tag{22}$$

Where $\mathrm{LCS}(X,Y)$ is the length of a longest common subsequence of $X$ and $Y$, and $\beta = P_{lcs}/R_{\mathrm{lcs}}$ when $\partial F_{lcs}/\partial R_{lcs} = \partial F_{lcs}/\partial P_{lcs}$. In DUC, $\beta$ is set to a very big number ($\rightarrow \infty$). Therefore, the LCS-based F-measure, *i.e.*, Equation 22, is Rouge-L.

**CIDEr.** The $\mathrm{CIDEr}_n$ (Vedantam et al., 2015) score for $n$-grams of length $n$ is computed using the average cosine similarity between the candidate sentence and the reference sentences, which accounts for both precision and recall:

$$CIDEr_n(c_i, S_i) = \frac{1}{m} \sum_j \frac{\mathbf{g^n}(c_i) \cdot \mathbf{g^n}(s_{ij})}{\|\mathbf{g^n}(c_i)\| \|\mathbf{g^n}(s_{ij})\|}, \tag{23}$$

where $\mathbf{g^n}(c_i)$ is a vector formed by $g_k(c_i)$ corresponding to all $n$-grams of length $n$ and $\|\mathbf{g^n}(c_i)\|$ is the magnitude of the vector $\mathbf{g^n}(c_i)$. Similarly for $\mathbf{g^n}(s_{ij})$.

Higher order (longer) $n$-grams are used to capture grammatical properties as well as richer semantics. (Vedantam et al., 2015) combine the scores from $n$-grams of varying lengths as follows:

$$CIDEr(c_i, S_i) = \sum_{n=1}^{N} w_n CIDEr_n(c_i, S_i), \tag{24}$$

Empirically, Vedantam et al.(Vedantam et al., 2015) found that uniform weights $w_n = 1/N$ work the best. So, we also use $N$ = 4.

**Dist-2.** Li et al. (2015) report the degree of diversity by calculating the number of distinct unigrams and bigrams in generated responses. The value is scaled by the total number of generated tokens to avoid favoring long sentences:

$$Dist(n) = \frac{Count(unique\ n-gram)}{Count(n-gram)} \tag{25}$$

### A.5 Dataset Details

**DailyDialog** consists of dyadic conversations covering various topics of daily-life and is designed to reflect natural human communication. Each utterance is annotated with both emotion categories and dialogue acts. The emotion labels include anger, disgust, fear, happiness, sadness, surprise, and neutral.

**MELD** is an extension of the EmotionLines dataset and is a multimodal corpus collected from the TV series *Friends*. It contains over 1,400 dialogues and 13,000 utterances, where each utterance is annotated with both emotion and sentiment labels.

**EmoryNLP** is also derived from the TV series *Friends*, but differs from MELD in the selection of scenes and the definition of emotion categories. It consists of multi-party conversations with utterances annotated into seven emotion classes.

**IEMOCAP** is a widely used benchmark dataset in affective computing, consisting of approximately 12 hours of audiovisual recordings collected from dyadic interactions. Each conversation is segmented into utterances annotated with both categorical emotion labels (e.g., anger, happiness, sadness, and neutral) and continuous Valence-Arousal values.

### A.6 Principle of Human Scoring

We start with the criteria proposed by Kang et al. (2024). The human evaluation is aimed to align with the ultimate purpose of emotional dialogue, the seeker's *satisfaction*. To achieve this, the supporter's behavior can be further classified into the following criteria:

*Acceptance*: Does the seeker accept without discomfort;

*Effectiveness*: Is it helpful in shifting negative emotions or attitudes towards a positive direction;

*Sensitivity*: Does it take into consideration the general state of the seeker. Furthermore, to clarify the capability of LLMs to align strategy and responses, we include Alignment.

To achieve a more elaborate assessment, we consider three more dimensions addressing the generation quality:

*Fluency*: the level of fluency of response.

*Emotion*: the emotional intensity of response which could affect the seeker's emotional state.

*Interesting*: Whether the response can arouse the seeker's interest and curiosity, presenting unique

ideas, vivid expressions or engaging elements that capture the seeker's attention and make the interaction more appealing.

We engage our interns as human evaluators to rate the models according to these multiple aspects, namely Fluency, Emotion, Interesting, and Satisfaction, with Satisfaction covering Acceptance, Effective, Sensitivity, and Satisfaction itself.
Throughout this evaluation process, we strictly comply with international regulations and ethical norms, ensuring that all practices conform to the necessary guidelines regarding participant involvement and data integrity.
Evaluators are required to independently evaluate each sample in strict accordance with the pre - established criteria. By adhering to these principles, the evaluation process maintains objectivity, standardization, and consistency, thus enhancing the overall quality and credibility of the evaluation results.
The detailed manual scoring criteria are as follows:

- Fluency:

  1: The sentence is highly incoherent, making it extremely difficult to understand and failing to convey a meaningful idea.

  2: The sentence has significant incoherence issues, with only parts of it making sense and struggling to form a complete thought.

  3: The sentence contains some incoherence and occasional errors, but can still convey the general meaning to a certain extent.

  4: The sentence is mostly fluent with only minor errors or slight awkwardness in expression, and effectively communicates the intended meaning.

  5: Perfect. The sentence is completely fluent, free of any errors in grammar, punctuation, or expression, and clearly conveys the idea.

- Emotion:

  1: The emotional expression is extremely inappropriate and chaotic, not in line with the content, and may convey wrong emotions.

  2: The emotional expression has obvious flaws, either too weak or exaggerated, and is disjointed from the content.

  3: The emotional expression is average. It can convey basic emotions but lacks depth and has minor issues.

  4: The emotional expression is good. It can effectively convey the intended emotion with an appropriate intensity and is well integrated with the content.

  5: The emotional expression is excellent. It is rich, nuanced, and perfectly matches the content, capable of evoking a strong and appropriate emotional response.

- Acceptance:

  1: The response inescapably triggers emotional resistance.

  2: The response is highly likely to trigger emotional resistance.

  3: The response has a possibility of emotional resistance occurring.

  4: The response rarely provokes emotional resistance.

  5: The response has no occurrence of emotional resistance.

- Effectiveness:

  1: The response actually worsens the seeker's emotional distress.

  2: The response carries the risk of increasing stress levels, and this outcome varies depending on the individual user.

  3: The response fails to alter the seeker's current emotional intensity and keeps it at the same level.

  4: The response shows promise in calming the emotional intensity; however, it is overly complicated or ambiguous for the user to fully comprehend and utilize effectively.

  5: The response appears to be highly effective in soothing the seeker's emotions and offers valuable and practical emotional support.

- Sensitivity:

  1: The response renders inaccurate evaluations regarding the seeker's state.

  2: The response is characterized by rash judgments, as it lacks adequate assessment and in-depth exploration of the seeker's state.

  3: The response is formulated with a one-sided judgment and a limited exploration of the seeker's state.

4: The response demonstrates an understanding that only covers a part of the seeker's state.

5: The response precisely grasps the seeker's state and is appropriately tailored according to the seeker's actual situation.

- Alignment:

  1: The response is in total contradiction to the predicted strategy.

  2: The response has a minor deviation from the predicted strategy.

  3: There is some ambiguity between the response and the predicted strategy.

  4: The response largely matches the predicted strategy, yet it contains some ambiguous elements.

  5: The response effectively makes itself consistent with the predicted strategy.

- Satisfaction:

  1: The response is extremely disappointing. It doesn't answer the question at all and is of no help.

  2: The response is poor. It only gives a partial answer and leaves many doubts unresolved.

  3: The response is average. It meets the basic requirements but isn't particularly outstanding.

  4: The response is good. It answers the question clearly and provides some useful details.

  5: The response is excellent. It not only answers the question perfectly but also offers valuable additional insights.

# B More Experimental Results

## B.1 Inter-Annotator Consistence

To indicate the scoring consistences of different human annotators, we analyze the their Fluency scoring statistics and present the Cohen's kappa matrix in Table 10. It can be observed that most $\kappa$ values fall between 0.8 and 1, indicating high inter-annotator agreement.

## B.2 Human Verification on Plutchik Scores

To further validate the reliability of our GPT-based rewarding mechanism (the Plutchik score $r_{\text{Plu}}$), we conduct additional human annotations on such rewards, and and validate their consistences. To achieve this, we randomly select 100 samples, then ask the volunteers to annotate the reward according to the **Plutchik's Wheel of Emotion** theory, on the same reward scale and dimensions: Alignment, Transition, and Function. We report the means and standard deviations of human and GPT-4o scores, as well as the correlation coefficient $\rho$ and Cohen's kappa $\kappa$ between them, to indicate their correlation levels.

| | eval_0 | eval_1 | eval_2 | eval_3 | eval_4 | eval_5 | eval_6 | eval_7 | eval_8 | eval_9 | eval_9 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **eval_0** | 1.00 | 0.83 | 0.81 | 0.76 | 0.82 | 0.81 | 0.80 | 0.81 | 0.81 | 0.82 | 0.82 |
| **eval_1** | 0.83 | 1.00 | 0.83 | 0.78 | 0.85 | 0.83 | 0.83 | 0.83 | 0.84 | 0.84 | 0.84 |
| **eval_2** | 0.81 | 0.83 | 1.00 | 0.78 | 0.83 | 0.82 | 0.82 | 0.82 | 0.83 | 0.82 | 0.82 |
| **eval_3** | 0.76 | 0.78 | 0.78 | 1.00 | 0.78 | 0.77 | 0.76 | 0.77 | 0.78 | 0.78 | 0.78 |
| **eval_4** | 0.82 | 0.85 | 0.83 | 0.78 | 1.00 | 0.83 | 0.83 | 0.83 | 0.83 | 0.82 | 0.82 |
| **eval_5** | 0.81 | 0.83 | 0.82 | 0.77 | 0.83 | 1.00 | 0.81 | 0.83 | 0.83 | 0.82 | 0.82 |
| **eval_6** | 0.80 | 0.83 | 0.82 | 0.76 | 0.83 | 0.81 | 1.00 | 0.81 | 0.81 | 0.81 | 0.81 |
| **eval_7** | 0.81 | 0.83 | 0.82 | 0.77 | 0.83 | 0.83 | 0.81 | 1.00 | 0.82 | 0.83 | 0.83 |
| **eval_8** | 0.81 | 0.84 | 0.83 | 0.78 | 0.83 | 0.83 | 0.81 | 0.82 | 1.00 | 0.82 | 0.82 |
| **eval_9** | 0.82 | 0.84 | 0.82 | 0.78 | 0.82 | 0.82 | 0.81 | 0.83 | 0.82 | 1.00 | 1.00 |

Table 10: the Cohen's Kappa Matrix among Evaluators of Fluency Score

| | mean | std | $\rho$ | $\kappa$ |
|---|---|---|---|---|
| Alignment | | | | |
| GPT-4o | 4.31 | 1.00 | 0.87 | 0.58 |
| Human | 4.16 | 0.94 | - | - |
| Emotion | | | | |
| GPT-4o | 4.13 | 1.16 | 0.96 | 0.78 |
| Human | 4.26 | 1.08 | - | - |
| Effectiveness | | | | |
| GPT-4o | 3.86 | 1.33 | 0.95 | 0.72 |
| Human | 3.69 | 1.21 | - | - |

Table 11: Statistical comparison of $r_{Plu}$ between human and GPT

Table 11 exhibits this experiment results. While the GPT-4o scores differ slightly from the human scores in terms of mean values, the results of $\rho$ and $\kappa$ indicate a high level of agreement between these two rewarding methods.

## B.3 Sensitivity Study on $w$

The weight of Plutchik score, $w$ is an important hyper-parameter, where higher values of $w$ emphasize theory-driven emotional consistency, whereas lower values rely more heavily on dataset annotations. Therefore, we conduct a sensitivity study on it, with results shown in Table 12. Results show that our framework is flexible and robust to different settings of $w$, while the optimal balance differs across datasets. This dataset-dependent characteristic suggests that different corpora exhibit varying degrees of annotation noise and emotional ambiguity, and our reward formulation allows this trade-off to be adjusted accordingly.

| $w$ | 1 | 0.9 | 0.8 | 0.7 | 0.5 | 0.3 | 0 |
|---|---|---|---|---|---|---|---|
| DailyDialog | 0.78 | 0.73 | 0.69 | 0.65 | 0.57 | 0.48 | 0.35 |
| EmoryNLP | 0.53 | 0.58 | 0.61 | 0.65 | 0.71 | 0.79 | 0.88 |
| MELD | 0.82 | 0.82 | 0.81 | 0.81 | 0.86 | 0.91 | 0.85 |
| IMEOCAP | 0.53 | 0.58 | 0.61 | 0.67 | 0.75 | 0.83 | 0.88 |

Table 12: Averaged rewards obtained on different values of $w$.

### B.4 The Self-EmoQ Algorithm

We illustrate the training mechanism of Self-EmoQ by the pseudo-codes in Algorithm 1.

**Algorithm 1** Self-EmoQ

Initialize the batch sizes $B$
Initialize replay buffer $\mathcal{B}$ with original dataset, Q $Q_\theta(s,a)$, target Q $\hat{Q}_{\theta^-}(s,a)$ with $\theta^- \leftarrow \theta$
Load pretrained model $g$, Set exploration rate $\epsilon$, discount factor $\gamma$, and update interval $C$, weight of Plutchik score $w$.
**while** not converged **do**
    Draw $B$ data $\{(h_t, x_t^u)\}$ from $\mathcal{B}$
    **for** each data **do**
        Form state $s_t = (h_t, s_t^u)$
        Select $e_t^s$ with $\epsilon$-greedy policy by Eq(1)
        Obtain $x_t^s$ , $r_t$ by Eq(10)
        Update history $h_{t+1}$; state $s_{t+1}$
        Store transition $(s_t, a_t, r_t, s_{t+1})$ in $\mathcal{B}$
    **end for**
    Sample batch $\{(s_i, a_i, r_i, s_i')\}$ from $\mathcal{B}$
    Compute TD target:

$$y_i = r_i + \gamma \max_{a'} \hat{Q}_{\theta^-}(s_i', a')$$

    Update Q-network by minimizing Eq.(3)
    Every $C$ steps update $\hat{Q}$ by $\theta^- \leftarrow \theta$
**end while**

### B.5 More Typical Cases

Besides the good case of IMEOCAP shown in Table 8, Table 13 further provides more typical cases on DailyDialog, EmoryNLP, and MELD.

| | |
|---|---|
| **Case of DailyDialog** | |
| History | user: (*no emotion*) Hi , Monica , how is everything going.<br>assistant: (*no emotion*) Everything goes well , but I am thinking about quitting my current job.<br>user: (*surprise*) Why. You ' re not satisfied anymore.<br>assistant: (*no emotion*) I just sense. But I cannot grow anymore. My boss is not really supporting me. I am interested in some positions in other JV companies , but I need to do some more in-depth research before I send my application letters out.<br>user: (*happiness*) That is important. Doing research on a company you are interested in will definitely help your application. |
| 0-shot | (*surprise*)That's a good point, I didn't think about the importance of researching the company before applying. |
| FSM | (*no emotion*) Thank you for your advice. |
| **Self-EmoQ** | (*happiness*) I will also need to update my resume. I've heard that the most important thing in a resume is to make it short and concise. |
| **Case of EmoryNlP** | |
| History | Joey Tribbiani: (*Neutral*) Hello.<br>Chandler Bing: (*Scared*) Hey, it's me. I know you can't stand to be in the same room as me, so I just thought I'd try and apologize over the phone. All I...<br>Joey Tribbiani: (*Neutral*) Hello.<br>Chandler Bing: (*Mad*) Look I never should have kissed your girlfriend, but I'm... |
| 0-shot | (*Anger*)Oh, so now you're apologizing? You think a simple phone call is going to make everything okay? You kissed my girlfriend, Chandler. That's not something you just apologize for over the phone. |
| FSM | (*Neutral*) I don't want to listen to it. |
| **Self-EmoQ** | (*Anger*) I don't want to hear it. |
| **Case of MELD** | |
| History | Joey: (*sadness*) God, it's gonna so weird like when I come home and you're not here.<br>Joey: (*sadness*) No more Joey and Chan's.<br>Joey: (*sadness*) No more J and C's.<br>Joey: (*neutral*) "You wanna go over to Joey and Chandler's?"<br>Joey: (*neutral*) "Can't, its not there."<br>Chandler: (*neutral*) Look, I'm just gonna be across the hall, we can still do all the same stuff |
| 0-shot | (*Neutral*) Oh, okay, that sounds cool, we can still hang out. |
| FSM | (*sadness*) No we can't! |
| **Self-EmoQ** | (*sadness*) Yeah, but it's not gonna be the same. |

Table 13: More Good Cases of Self-EmoQ.