Third-Person Appraisal Agent: Simulating Human Emotional Reasoning in Text with Large Language Models

Anonymous ACL submission

Abstract

Emotional reasoning is essential for improving human-AI interactions, particularly in mental health support and empathetic systems. However, current approaches, which primarily map sensory inputs to fixed emotion labels, fail to capture the intricate relationships between motivations, thoughts, and emotions, thereby limiting their ability to generalize across diverse emotional reasoning tasks. To address this, we propose a novel third-person appraisal agent that simulates human-like emotional reasoning through three phases: Primary Appraisal, Secondary Appraisal, and Reappraisal. In the Primary Appraisal phase, a third-person generator powered by a large language model (LLM) infers emotions based on cognitive appraisal theory. The Secondary Appraisal phase uses an evaluator LLM to provide feedback, guiding the generator in refining its predictions. The generator then uses counterfactual reasoning to adjust its process and explore alternative emotional responses. The Reappraisal phase utilizes reinforced fine-tuning (ReFT) by employing a reflective actor-critic framework to further enhance the model's performance and generalization. This process uses reward signals and learns from appraisal trajectories without human annotations. Our approach outperforms baseline LLMs in various emotion reasoning tasks, demonstrating superior generalization and interpretability. To the best of our knowledge, this is the first cognition-based architecture designed to enhance emotional reasoning in LLMs, advancing AI towards human-like emotional understanding. The code is available here.

1 Introduction

Emotional reasoning is a critical cognitive process focused on understanding and interpreting emotions by analyzing the intricate relationships between a speaker's motivations, thoughts, and emotional expressions. This capability is essential in fields such as mental health support systems and empathetic conversational AI, as enhancing a model's ability to comprehend human emotions can significantly advance human-AI interaction. However, existing studies (Wondra and Ellsworth, 2015; Ribeiro et al., 2016; Hazarika et al., 2018; Ong et al., 2019; Jiao et al., 2020; Vellido, 2020; Gao et al., 2021; Hu et al., 2021a; Li et al., 2022; Sabour et al., 2022; Zhao et al., 2022; Cortiñas-Lorenzo and Lacey, 2023; Hu et al., 2023) primarily focus on feature extraction-based approaches that map sensory inputs to a fixed set of emotion labels, which limits the model's ability to generalize across diverse emotion reasoning tasks. To address this, emotional analysis must evolve beyond static labels and adopt human-like cognitive reasoning, establishing connections between emotions and their underlying causes. This leads to a critical research question: How can we develop emotion reasoning approaches that more closely mimic human understanding of emotions in various contexts?

The appraisal theory of emotion (Lagattuta et al., 1997; Wondra and Ellsworth, 2015; Ong et al., 2019) posits that emotions arise from individuals' appraisals (i.e., cognitive evaluations) of situations, particularly in relation to their goals, desires, intentions, or expectations. Inspired by this theory, we developed an agentic workflow called the "third-person appraisal agent"(see figure 1a) to simulate three stages of the human cognitive appraisal process (Roseman and Smith, 2001; Ellsworth and Scherer, 2003; Watson and Spence, 2007): primary appraisal, secondary appraisal, and reappraisal. The goal is to enable the agent to evaluate and understand emotions in a manner that closely resembles how humans process emotions.

In the primary appraisal phase, we design an LLM, termed the third-person appraisal generator LLM, which acts as an external observer. This model first analyzes conversations to evaluate how

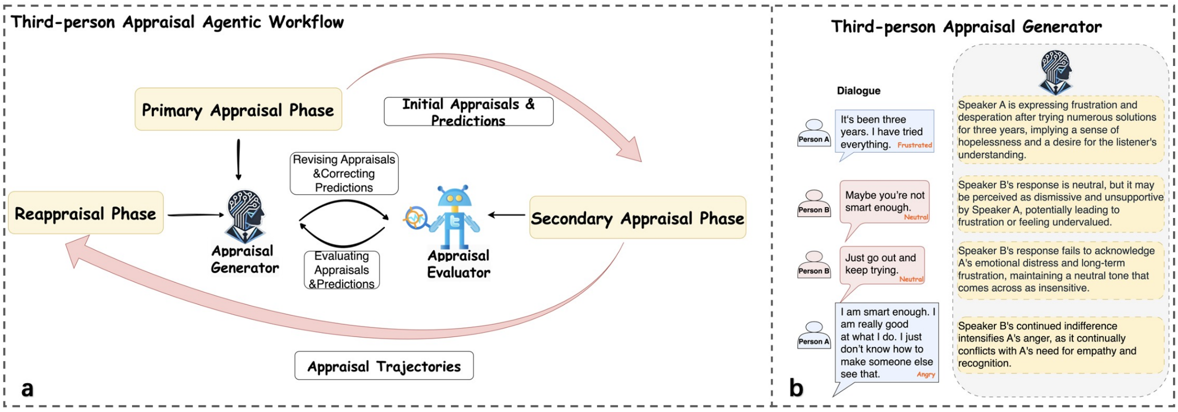

Figure 1: a: An overview of the Third-Person Appraisal Agentic Workflow used to fine-tune the Third-Person Appraisal Generator. b: Performance of the fine-tuned Third-Person Appraisal Generator on conversational emotion analysis. The example sample is drawn from the IEMOCAP dataset (Busso et al., 2008).

contextual utterances align with an interlocutor's objectives and expectations, subsequently inferring emotional predictions. For example, as shown in Figure 1b, Person A's anger may arise from Person B's indifferent attitude, which contradicts Person A's expectations. By simulating the cognitive appraisal process, the third-person appraisal generator can better interpret emotional dynamics within conversational contexts.

We regard secondary appraisal as a reflective process that follows primary appraisal. When the initial evaluation is determined to be inaccurate, the agent adjusts its appraisals based on the identified errors and subsequently updates its emotional predictions. To simulate this process, we introduce an additional LLM, termed the Appraisal Evaluator LLM, which evaluates the performance of the Appraisal Generator LLM. The Appraisal Generator LLM refines its emotional appraisals through counterfactual reasoning (Roese, 1997) by hypothesizing alternative emotional responses and adjusting its appraisal process based on how well these alternatives align with contextual factors. In this way, the secondary appraisal process enables the Appraisal Generator LLM to refine its reasoning steps and improve its predictions using feedback from the Appraisal Evaluator LLM. This entire phase is implemented within a verbal reinforcement learning (RL)-based framework(Shinn et al., 2024), where the agent continuously generates and refines appraisals through an iterative reflective loop that gathers reflection samples.

Although the secondary appraisal phase yields more accurate appraisals, further fine-tuning is necessary to enhance the model's generalization capabilities. To address this, we introduce a reappraisal phase to further refine the model using the reinforced fine-tuning (ReFT) framework (Trung et al., 2024). Specifically, we employ a reflective actor-critic reinforcement learning method (Flavell et al., 2001; Haarnoja et al., 2018) in this work. During the reappraisal phase, ReFT integrates reward signals into the model's learning process, refining its performance by learning from appraisal trajectories collected during the secondary appraisal phase — all without the need for human annotations. To the best of our knowledge, this work is among the first to incorporate a ReFT-based method to improve the emotional reasoning capabilities of LLMs.

Meanwhile, the efficient and reproducible evaluation of emotion reasoning remains challenging due to the reliance on manual annotations(Kazienko et al., 2023; Madaan et al., 2024; Huang et al., 2024), which are time-consuming, costly, and highly variable. This variability limits large-scale model comparisons and hinders the reliable replication of results. We aim to simplify emotion reasoning performance evaluation by enabling LLMs to automatically assess and score emotional reasoning tasks. Specifically, we evaluate: (1) Emotional Comprehension, which assesses the ability to recognize emotional causes and understand the speaker's motivations;(2) Contextual Understanding, which measures the understanding of context and how emotions evolve within a conversation; and (3) Expressive Coherence and Performance, which evaluates whether the model communicates its emotional reasoning clearly and is easy to understand. Based on these three evaluation criteria, we have developed a six-dimensional evaluation system. By transforming this system into a multiple-choice format, we enable LLMs to evaluate emotional reasoning tasks efficiently and reproducibly.

Our experiments demonstrate the effectiveness of our approach, as it outperforms LLM baselines in accuracy across various emotional reasoning tasks, including reasoning about previously unseen emotions from new conversational contexts and vicarious emotions such as empathy and distress from written essays. The main contributions of this paper are summarized as follows:

- We introduce a novel third-person appraisal agent designed to simulate human-like cognitive reasoning for emotions. To our knowledge, this is the first attempt to enhance the emotional reasoning abilities of LLMs by guiding them to evaluate emotions through the lens of cognitive appraisal theory.
- To enhance the reasoning capabilities and generalization performance of the third-person appraisal generator LLM, we incorporate both the secondary appraisal and reappraisal phases into the agentic workflow. First, the model uses counterfactual thinking to generate reflections. Second, it employs a reflective actor-critic RL strategy to fine-tune its reasoning capabilities by leveraging these reflections as a limited set of demonstration examples. Experimental results show significant improvements in prediction accuracy and generalization across various emotional reasoning tasks.
- We also develop a six-dimensional comprehensive evaluation system to assess the emotional reasoning capabilities of LLMs, which, to the best of our knowledge, is the first such attempt in this field. This evaluation offers a reproducible, explainable, and efficient alternative to traditional manual annotations.

2 Related Work

Self-Reflection: Current approaches to emotion reasoning with LLMs emphasize prompt tuning for tasks such as emotional cause extraction (Doe and Smith, 2023; Bhaumik and Strzalkowski, 2024; Belikova and Kosenko, 2024). However, there is limited research exploring the integration of self-reflection or feedback mechanisms specifically within emotion reasoning tasks. Currently, self-reflection or feedback mechanisms have been explored in other domains, such as mathematical reasoning, code generation, and so on (Welleck et al., 2022; Yang et al., 2022; Paul et al., 2023; Madaan et al., 2024; Shinn et al., 2024). Shinn et al. (2024) introduces Reflexion, a self-reflection mechanism that enables LLMs to improve their reasoning capabilities by learning from past mistakes. However, the application of Reflexion to emotion reasoning tasks has yet to be thoroughly investigated. Although Madaan et al. (2024) demonstrates the effectiveness of self-reflection in sentiment style transfer—a task that modifies a text's sentiment while preserving its meaning—this task is only tangentially related to emotion reasoning. In contrast, our work uniquely combines counterfactual reasoning with a reflection mechanism. Our framework not only enables LLMs to generate self-feedback and refine their predictions, but also aligns them with human-like emotion reasoning processes, thereby simulating how humans understand emotions.

Reinforcement Learning In our task, we employ an actor-critic reinforcement learning framework to align AI systems with human preferences (Ouyang et al., 2022). Recently, several novel training algorithms have emerged to enhance alignment effectiveness, including Proximal Policy Optimization (PPO) (Schulman et al., 2017), Direct Preference Optimization (DPO) (Rafailov et al., 2024), Identity Preference Optimization (IPO) (Azar et al., 2024), and Kahneman-Tversky Optimization (KTO) (Ethayarajh et al., 2023). While these methods primarily focus on enhancing alignment, our approach goes a step further by uniquely integrating reinforcement learning as a fine-tuning paradigm (ReFT) to improve emotional reasoning. This process mirrors how humans iteratively refine their thought processes to achieve a deeper understanding of emotions, ultimately yielding superior performance compared to conventional supervised fine-tuning.

3 Problem Description: Formulating as a Generative Task

We propose a generative approach for zero-shot emotion reasoning and prediction based on textual input. Each utterance is associated with a specific speaker and a set of emotional categories that may vary depending on the emotional types present in different dialogue datasets. The objectives are twofold: (1) generate an appraisal a_i for each utterance u_i, and (2) infer an emotion label $\hat{y}_i$ based on this appraisal. For contextual understanding, we define a window of length l to gather the dialog context C_i for each utterance u_i. This context consists of the current utterance and the preceding $l-1$

utterances, along with their corresponding speaker information.

Therefore, we frame the generative task in a **question-and-answer** format, as illustrated below:

> **Question:** Given the dialogue context C_i, predict an emotion label for the target utterance u_i. Choose from *happy, sad, neutral, angry, excited, frustrated*.
> **Answer:** Generate an appraisal a_i for the target utterance u_i and then produce the final prediction $\hat{y}_i$.

3.1 Three cognitive appraisal phases for Third-person appraisal agent

We introduce a third-person appraisal agent composed of two specialized LLMs: the Appraisal Generator and the Appraisal Evaluator. This agentic workflow consists of three phases—primary appraisal, secondary appraisal, and reappraisal—which together enable the model to simulate human cognitive appraisal from a third-person perspective (see Figure 1a).

Appraisal Generator LLM: The appraisal generator M_A is responsible for generating appraisals and making predictions based on those appraisals. We prompt M_A with an AppraisalInstruction prompt (see Appendix C) to generate an appraisal a_i and a predicted emotion label $\hat{y}_i$, given only the input utterance u_i and its corresponding dialogue context C_i.

Appraisal Evaluator LLM: The Evaluator M_E assesses the accuracy of the appraisals and predictions, providing feedback upon which reward values are assigned. We utilize M_E to provide two types of rewards:

- **Action Reward** r^{actor}: Assigns 0 for correct emotion label predictions and -1 for incorrect ones, reinforcing accurate predictions and guiding the model to refine its appraisals.

- **Critic Reward** r^{critic}: Evaluates the alignment of each appraisal's valence-arousal (VA) vector with its target emotion class. Valence and arousal scores are obtained from the NRC-VAD lexicon (Mohammad, 2018) and normalized to the range [-1, 1] using min-max scaling. In the Evaluation Prompt (see Appendix C), the Evaluator M_E uses the Circumplex Model (Russell, 1980) to classify emotion labels into predefined valence and arousal ranges. It then checks if the appraisals fall within these ranges, assigning a score of 0 for alignment and -1 for misalignment.

Algorithm 1 VRL: Secondary Appraisal via Counterfactual Reasoning

Require: Input u_i, dialog context C_i, models $\{M_A, M_E\}$, prompts $\{p_a, p_c\}$, true emotion label y_i

```
(a_{i,0}, ŷ_{i,0}) = M_A(p_a ‖ u_i ‖ C_i)                    ▷ Initial generation (Eq.1)
(r^{actor}_{i,0}, r^{critic}_{i,0}) = M_E(ŷ_{i,0}, y_i, a_{i,0})   ▷ Initial feedback (Eq.2)
Add (u_i, a_{i,0}, r^{actor}_{i,0}, r^{critic}_{i,0}) to appraisal trajectory D_i
for iteration k = 1, 2, ... do
    (a_{i,k}, ŷ_{i,k}) = M_A(p_c^k ‖ u_i ‖ x_i ‖ {ŷ_{i,0}, ..., ŷ_{i,k-1}})   ▷ Counterfactual reasoning (Eq.3)
    (r^{actor}_{i,k}, r^{critic}_{i,k}) = M_E(ŷ_{i,k}, y_i, a_{i,k})   ▷ Feedback (Eq.4)
    Add (u_i, a_{i,k}, r^{actor}_{i,k}, r^{critic}_{i,k}) to appraisal trajectory D_i
    if ŷ_{i,k} = y_i then                                   ▷ Stop condition
        break
    end if
end for
return D_i
```

3.1.1 Primary Appraisal Phase

We prompt the LLM $M_A$using AppraisalInstruction Prompt p_a, which is designed based on the principles of cognitive appraisal theory (Watson and Spence, 2007; Ong et al., 2019), to generate an appraisal a_i for utterance u_i. This process can be formulated as:

$$(a_{i,0},\ \hat{y}_{i,0}) = M_A(p_a \| u_i \| C_i) \tag{1}$$

The goal of generating appraisals is to enable the model to reason about emotions by evaluating how each participant's goals, desires, intentions, and expectations align with the conversational context. To accomplish this, we introduce a primary appraisal phase in which the model learns to generate appraisals from a third-person perspective, thereby enhancing its capacity to analyze emotional dynamics through a cognitive process.

3.1.2 Secondary Appraisal Phase

The secondary appraisal process utilizes the appraisal generator M_A to create new appraisals by adjusting its previous ones based on feedback from M_E (see Algorithm 1).

The secondary appraisal framework is detailed in the following steps: We first evaluate the initial appraisal and prediction generated from the primary appraisal phase with appraisal evaluator LLM,M_E, obtaining actor and critic rewards:

$$(r^{\text{actor}}_{i,0},\ r^{\text{critic}}_{i,0}) = M_E(\hat{y}_{i,0},\ y_i,\ a_{i,0}) \tag{2}$$

If the initial prediction $\hat{y}_{i,0}$ is incorrect, M_A enters an iterative counterfactual reasoning loop to generate new appraisals. At each iteration k ($k \geq 1$), the CounterfactualReasoning p_c^k (see Appendix C) uses the history of incorrect predictions $\{\hat{y}_{i,0}, \hat{y}_{i,1}, \ldots, \hat{y}_{i,k-1}\}$ to update the output for utterance u_i:

$$(a_{i,k},\ \hat{y}_{i,k}) = M_A(p_c^k \| u_i \| \{\hat{y}_{i,0}, \ldots, \hat{y}_{i,k-1}\}) \tag{3}$$

We then evaluate the updated appraisal with M_E:

$$(r_{i,k}^{\text{actor}},\ r_{i,k}^{\text{critic}}) = M_E(\hat{y}_{i,k},\ y_i,\ a_{i,k}) \tag{4}$$

This reflective process continues until the prediction is correct or a maximum number of iterations K is reached. After completing the secondary appraisal phase, we collect the appraisal trajectories into a replay buffer D:

$$D = \left\{ (u_{i,k}, a_{i,k}, r_{i,k}^{\text{actor}}, r_{i,k}^{\text{critic}}) \mid \begin{array}{c} k = 0, \ldots, K_i; \\ i = 1, \ldots, I \end{array} \right\}$$

K_i is the number of iterations for the i-th utterance. If M_A makes a correct prediction at $k = 0$, we set $K_i = 0$, and the trajectory consists only of the initial appraisal.

Algorithm 2 ReFT: Reappraisal via Reflective Actor-Critic RL

Initialize Appraisal generator, Critics Q_{θ_1} and Q_{θ_2}, Value Function V_ψ, and Replay Buffer $\mathcal{D}$ (an offline dataset).
Initialize Policy $\pi_\phi(a_{i,k'}|u_{i,k'})$, where ϕ is the set of parameters of the appraisal generator.
Set $t \leftarrow 0$
while $t <$ T **do**
  Sample batch $\{u_{i,k'}, a_{i,k'}, r_{i,k'}, a_{i,k'+1}\}$ from $\mathcal{D}$.
  For terminal steps (where $k' = K_i$), **set** $a_{i,k'+1} = a_{i,k'}$.
  Critic Update: Minimize J_Q for Q_{θ_1} and Q_{θ_2} (Eq.6)
  Value Function Update: Minimize J_V (Eq.7)
  Update target networks $Q_{\bar{\theta}_1}$, $Q_{\bar{\theta}_2}$, and $V_{\bar{\psi}}$ via Polyak averaging
  Compute Advantage: $A(u_{i,k'}, a_{i,k'})$ (Eq.8)
  Actor Update: Minimize J_ϕ (Eq.9)
  Increment $t \leftarrow t + 1$
end while
return Appraisal Mechanism π_ϕ

3.1.3 Reappraisal Phase

The Reappraisal Phase enhances the model's emotional reasoning through reward-based learning. After generating appraisal trajectories in the Secondary Appraisal Phase, the model enters the Reappraisal Phase, where it fine-tunes its predictions using a ReFT framework(see Algorithm 2). This phase employs a reflective actor-critic method: the Actor (appraisal generator) proposes appraisals, while the Critic evaluates the Actor's performance and provides feedback. The iterative interaction between the Actor and Critic continuously refines the Actor's appraisal mechanism, thereby improving its reasoning capability.

We use off-policy learning, allowing the Critic to learn from a broader set of experiences by sampling from the replay buffer $\mathcal{D}$, which is obtained during the secondary appraisal phase. This approach improves stability and efficiency by leveraging past appraisals and rewards.

Critic Model: The Critic evaluates the appraisals and provides value estimates to guide the Actor's policy refinement. We train three Multi-Layer Perceptrons (MLPs)(Taud and Mas, 2018): two critics representing utterance-level Q-functions, $Q_{\theta_1}(u_{i,k'}, a_{i,k'})$ and $Q_{\theta_2}(u_{i,k'}, a_{i,k'})$, where $u_{i,k'}$ and $a_{i,k'}$ are sampled from $\mathcal{D}$. The double critic architecture is employed to reduce overestimation bias. Additionally, we have an MLP for the utterance-level value function $V_\psi(u_{i,k'})$. In this framework, k' represents the iteration index in $\mathcal{D}$ for the i-th utterance. It ranges from $k' = 0$ (initial appraisal) up to $k' = K_i$, where K_i is the total number of iterations for i-th utterance.

Target networks $Q_{\bar{\theta}_1}$ and $Q_{\bar{\theta}_2}$, and $V_{\bar{\psi}}$ are delayed copies of the respective models, updated via Polyak averaging (Polyak and Juditsky, 1992). The parameters θ_1, θ_2, and ψ are the trainable parameters of the MLPs, while the target network parameters $\bar{\theta}_1$, $\bar{\theta}_2$, and $\bar{\psi}$ are updated using the moving averages of θ_1, θ_2, and ψ, respectively.

The Q-functions are trained by minimizing the Bellman error using targets derived from $V_{\bar{\psi}}$. The value function V_ψ is trained to approximate the expected value of $Q_{\bar{\theta}_1}$ and $Q_{\bar{\theta}_2}$. To guide the appraisal process, we use a weighted combination of two reward signals:

$$r_{i,k'} = \alpha r_{i,k'}^{\text{actor}} + \beta r_{i,k'}^{\text{critic}} \tag{5}$$

$$\begin{aligned} J_Q(\theta_j) =& \mathbb{E}_{(u_{i,k'}, a_{i,k'}, r_{i,k'}) \sim \mathcal{D}} \Bigg[\Big(Q_{\theta_j}(u_{i,k'}, a_{i,k'}) \\ & - \big(r_{i,k'} + \gamma V_{\bar{\psi}}(u_{i,k'}) \big) \Big)^2 \Bigg], \quad j = 1, 2 \end{aligned} \tag{6}$$

$$\begin{aligned} J_V(\psi) =& \mathbb{E}_{(u_{i,k'}, a_{i,k'+1}) \sim \mathcal{D}} \Bigg[\big(V_\psi(u_{i,k'}) - Q_{\bar{\theta}_1}(u_{i,k'}, a_{i,k'+1}) \big)^2 \\ & + \big(V_\psi(u_{i,k'}) - Q_{\bar{\theta}_2}(u_{i,k'}, a_{i,k'+1}) \big)^2 \Bigg] \end{aligned} \tag{7}$$

where α and β are weighting coefficients, and γ is the discount factor. For terminal steps (i.e., when the process reaches its final step) where $k' = K_i$, we set $a_{i,k'+1} = a_{i,k'}$.

Actor Model: We train the appraisal generator using an offline policy gradient approach, utilizing advantage values derived from the minimum of the two Q-values from the critic model. The advantage function measures how much better a particular action (appraisal) is compared to the expected outcome, represented by the value function:

$$A(u_{i,k'}, a_{i,k'}) = \min\left(Q_{\theta_1}(u_{i,k'}, a_{i,k'}), Q_{\theta_2}(u_{i,k'}, a_{i,k'})\right) - V_\psi(u_{i,k'}) \quad (8)$$

These advantage values guide the M_A in refining its appraisal generation mechanism, leading to more accurate emotional appraisals. The policy gradient update is performed by minimizing:

$$J_\phi(\pi) = -\mathbb{E}_{(u_{i,k'}, a_{i,k'})\sim\mathcal{D}}\left[A(u_{i,k'}, a_{i,k'}) \log \pi_\phi(a_{i,k'}|u_{i,k'})\right] \quad (9)$$

where ϕ represents the trainable parameters of M_A.

4 Experiments & Results

In this section, we present three major experiments designed to evaluate the performance of our proposed model. The experiments are structured as follows: (1) a comparative analysis against LLM baseline models; (2) an ablation study assessing the agentic workflow (3) a qualitative analysis of the model's appraisal performance on the DailyDialog dataset. Additionally, we also present comparative analysis of two VRL-based strategies for evaluating the effectiveness of the secondary appraisal phase; details can be found in appendix B.

Baselines: For comparison, we use Mistral-7B-Instruct-v0.3, Gemma1.1-7B-Instruct, LLaMA3.1-8B-Instruct, and Mistral-Nemo-Instruct-2407-bnb-4bit as baseline models. Note: Mistral-Nemo-Instruct-2407 is a 12B-parameter LLM.

Evaluation Metrics: We report value accuracy for all three datasets, including IEMOCAP, DailyDialog and WASSA2023 datasets. For the WASSA2023 dataset, the accuracy for empathy and distress scores is computed based on the absolute difference between the predicted and gold-standard values, with a prediction considered correct if the absolute difference is less than or equal to 2.

Implementation Details: We set the fixed window length, l, to 5. The appraisal generator LLM, M_A, is implemented using the Mistral-7B-Instruct-v0.3 model, and the appraisal evaluator LLM, M_E, is implemented using the LLaMA3.1-8B-Instruct model. In the secondary appraisal phase, the reflective cycle is set to 5 iterations. During the reappraisal phase, each of the double critic models is implemented as a 3-layer MLP, while the value model is implemented as a 2-layer MLP, with their embeddings initialized using pre-trained RoBERTa (Liu, 2019). Both the actor and critic models are trained using the Adam optimizer (Kingma and Ba, 2014) with the same learning rate of 1×10^{-5}. Training is conducted over 10 epochs. The constant coefficients α and β are set to 0.9 and 0.45, respectively. The M_A model is trained using 4-bit quantized low-rank adapters (LoRA) (Hu et al., 2021b), with $r = 16$. During inference (the test mode), the model's temperature is set to 0.7.

The dataset information is provided in the appendix (see Appendix A).

4.1 Main Results

To evaluate the effectiveness of our third-person appraisal agent, we benchmark it against instruction-tuned LLM baselines. We select the first 1,000 utterances from the IEMOCAP training dataset, generating 2,204 appraisal trajectories during the secondary appraisal phase. These trajectories are then used to train the third-person appraisal generator in the reappraisal phase. Finally, the fine-tuned model is evaluated on the first 700 utterances from the IEMOCAP test set.

	Methods	Acc.
Zero-shot	[1] Mistral-7B-Instruct-v0.3	48.05
	[2] Gemma1.1-7B-Instruct	45.29
	[3] LLAMA-3.1-8B-Instruct (causal prompt)	40.13
	[4] LLAMA-3.1-8B-Instruct	46.75
	[5] Mistral-Nemo-12B-Instruct	50.18
SFT	[6] Mistral-7B-Instruct-v0.3	51.45
	[7] Gemma1.1-7B-Instruct	49.29
	[8] LLAMA-3.1-8B-Instruct	48.71
	[9] Mistral-Nemo-12B-Instruct	52.56
Ours	[10] Mistral-7B-Instruct-v0.3	**54.57**

Table 1: Performance comparisons in value accuracy of our model against baselines on the IEMOCAP test set.

Table 1 presents a performance comparison of our method against baseline models on the IEMOCAP test set, with accuracy values reported for both zero-shot and fine-tuned configurations. Our method [10] achieves the highest accuracy of

Model	DailyDialog					WASSA	
	ang	**sad**	**neu**	**hap**	**surp**	**emp**	**dis**
Original	0.33	0.73	0.34	0.60	0.50	0.68	0.69
Ours	**0.78**	**0.80**	**0.64**	**0.61**	**0.81**	**0.72**	**0.75**

Table 2: Performance comparison between the original LLM and our model on unseen datasets: DailyDialog and WASSA2023. Abbreviations: ang (angry), neu (neutral), hap (happy), surp (surprise), emp (empathy), dis (distress).

54.57%, significantly outperforming all baseline models.

In the zero-shot setting, we observe that [3] uses the causal prompt (see Appendix C) from (Team, 2024) to guide LLAMA-3.1-8B-Instruct in identifying emotion triggers and using them to infer emotions. However, this method reduces performance by 7.92% compared to [4], likely due to the model's difficulty in understanding the causal relationship between emotional triggers and the speaker's emotional responses. Models [1,2,4,5] use a general prompt to infer emotions solely based on the provided dialogue context.

We observe that all models benefit from SFT compared to zero-shot. However, our method outperforms all baseline models in the SFT setting, even when learning from a limited number of training samples. Most strikingly, our 7B-parameter model [10] outperforms the larger 12B-parameter LLM in both zero-shot [5] and SFT settings [9]. This result underscores the efficiency and effectiveness of our approach, allowing the smaller model to achieve superior performance compared to larger models.

Furthermore, we evaluate the model's general reasoning capabilities without fine-tuning by testing it on two previously unseen datasets: 1,000 utterances from the DailyDialog test set to predict five different emotions in conversational data, and 208 essays from the WASSA test set to measure its ability to predict empathy and distress in written text. As shown in Table 2, our approach consistently outperforms the original LLM integrated into Mistral-7B-Instruct-v0.3 across all tasks.

4.2 Ablation Study on Third-person Appraisal Agentic Workflow

This ablation study (see table 3) shows that including all three appraisal phases results in the highest accuracy (54.57%), while excluding any phase leads to a performance drop. Specifically, excluding the primary appraisal phase causes a significant decrease to 46.44%. This indicates the significant impact of the primary appraisal phase, as it serves as the foundational appraisal-driven principle for emotion analysis. Since the goal of the Secondary Appraisal phase is to generate additional appraisal trajectories for the model to learn from during the final Reappraisal phase, we observe that removing this phase resulted in a significant decline in the model's overall performance. This finding highlights the crucial role of the Secondary Appraisal phase in the entire workflow.

We further evaluate the reappraisal phase through an ablation study, where we remove specific components from the model: 1) no actor rewards during RL, 2) no critic rewards during RL, and 3) no appraisal instruction, where the agent is instruction-tuned without the AppraisalInstruction Prompt. Table 4 demonstrates that incorporating both actor and critic rewards enhances the agent's appraisal capabilities, indicating that this RL strategy can further enhance the agent's ability to generate accurate appraisals and predictions. Conversely, removing the AppraisalInstruction prompt results in a significant 4.71% drop in accuracy, indicating that the appraisal-based instruction plays a crucial role in guiding the model's reasoning process (Chung et al., 2024).

4.3 The Performance of Appraisals

We compare the appraisals generated by the same original LLM with those generated by our third-person appraisal generator on the same 1,000 utterances from the DailyDialog test set. Two key improvements are observed in this experiment. First, our model demonstrates advanced reasoning by evaluating the speaker's mental states—such as attitudes, goals, desires, and expectations—using contextual information. Examples in Appendix E demonstrate the comparative reasoning performance of our appraisal generator against the original LLM. Our model effectively identifies underlying causes, such as the speaker's motivations and intentions, going beyond basic emotional triggers. In contrast, the original LLM primarily focuses on identifying emotion triggers and provides limited reasoning based on surface-level cues and sentiments.

Our model shows an improved ability to generate qualitative appraisals, which is a challenging task for LLMs as it requires understanding how conversational utterances influence emotions. To

Third-person Appraisal Agentic Workflow			
Primary Appraisal Phase	**Secondary Appraisal Phase**	**Reappraisal Phase**	**Acc.**
✓	×	✓	48.52
×	✓	✓	46.44
✓	×	×	51.00
✓	✓	✓	**54.57**

Table 3: The performance of the agentic workflow is evaluated on the IEMOCAP test set. The table highlights how the inclusion or exclusion of three different appraisal phases influences the agent's performance in terms of accuracy.

Model Setting	**Accuracy**
Mistral-7B-Instruct + **Reappraisal Phase**	54.57
- w/o Actor Rewards	54.29
- w/o Critic Rewards	54.14
- w/o Appraisal Instruction	49.86

Table 4: Ablation study on reappraisal phase.

Metric	Original	**Ours**
Sentiment Awareness	4.67	**4.97**
Contextual Understanding	4.52	**4.60**
Sensitivity to Emotional Causes	4.43	**4.88**
Emotional Dynamics Responsiveness	4.23	**4.32**
Motivational Understanding	4.42	**5.13**
Clarity and Coherence Assessment	4.55	**4.75**

Table 5: Comparison of appraisal quality between the original LLM and our third-person appraisal generator LLM.

assess our agent's appraisal quality compared to the original LLM, we develop a set of appraisal quality metrics and use GPT-4 to rate each appraisal on a scale of 1 to 6 using the same DailyDialog test set. The average scores for each metric are shown in Table 5, with detailed explanations provided in Appendix D. Based on these results, we make the following observations:

- The original LLM achieves the highest sentiment awareness score across all of its metrics, highlighting its strong emphasis on sentiment analysis in its reasoning process.
- Both models perform well on clarity and coherence, indicating their ability to generate well-structured appraisals.
- Our model excels in motivational understanding, demonstrating a strong focus on identifying motivations when analyzing emotions.
- Key metrics for evaluating the model's reasoning performance include sentiment awareness, contextual understanding, responsiveness to emotional dynamics, and comprehension of motivations. The table shows that our model outperforms the baseline model in all four metrics, demonstrating its superior reasoning capabilities for conversational emotion analysis.

5 Conclusion

We introduce a novel agentic workflow that enables the training of a model capable of enhancing emotional reasoning capabilities without human annotations. Specifically, this workflow allows the model to iteratively refine its emotional reasoning through reinforcement learning, even with a limited number of demonstration samples. Our approach advances the development of explainable AI by training the model to perform emotion reasoning in a way that more closely aligns with human emotional understanding.

6 Limitations

A key limitation of our work is the inherent difficulty LLMs face in interpreting complex emotional transitions. For example, understanding how an extremely positive emotion like 'happiness' can shift into an extremely negative one like 'sadness' remains a major challenge. Addressing these limitations will be a primary focus of our future research as we aim to further improve the agent's ability to comprehend and reason through complex emotion shifts.

References

Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. 2024. A general theoretical paradigm to understand learning from human preferences. In *International Conference on Artificial Intelligence and Statistics*, pages 4447–4455. PMLR.

C Daniel Batson, Jim Fultz, and Patricia A Schoenrade. 1987. Distress and empathy: Two qualitatively distinct vicarious emotions with different motivational consequences. *Journal of personality*, 55(1):19–39.

Julia Belikova and Dmitrii Kosenko. 2024. Deeppavlov at semeval-2024 task 3: Multimodal large language models in emotion reasoning. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1747–1757.

Ankita Bhaumik and Tomek Strzalkowski. 2024. Towards a generative approach for emotion detection and reasoning. *arXiv preprint arXiv:2408.04906*.

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*, 42:335–359.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Karina Cortiñas-Lorenzo and Gerard Lacey. 2023. Toward explainable affective computing: A review. *IEEE Transactions on Neural Networks and Learning Systems*.

John Doe and Alice Smith. 2023. Causal inference in customer feedback analysis: A benchmarking approach with llms. *Proceedings of the 10th Conference on Natural Language Processing*, 34(1):123–135.

Phoebe C Ellsworth and Klaus R Scherer. 2003. Appraisal processes in emotion.

Kawin Ethayarajh, Winnie Xu, Dan Jurafsky, and Douwe Kiela. 2023. Human-centered loss functions (halos). Technical report, Technical report, Contextual AI.

John H Flavell, Eleanor R Flavell, and Frances L Green. 2001. Development of children's understanding of connections between thinking and feeling. *Psychological science*, 12(5):430–432.

Jun Gao, Yuhan Liu, Haolin Deng, Wei Wang, Yu Cao, Jiachen Du, and Ruifeng Xu. 2021. Improving empathetic response generation by recognizing emotion cause in conversations. In *Findings of the association for computational linguistics: EMNLP 2021*, pages 807–819.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. 2018. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pages 1861–1870. PMLR.

Md Rakibul Hasan, Md Zakir Hossain, Tom Gedeon, Susannah Soon, and Shafin Rahman. 2023. Curtin OCAI at WASSA 2023 empathy, emotion and personality shared task: Demographic-aware prediction using multiple transformers. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 536–541, Toronto, Canada. Association for Computational Linguistics.

Devamanyu Hazarika, Soujanya Poria, Rada Mihalcea, Erik Cambria, and Roger Zimmermann. 2018. Icon: Interactive conversational memory network for multimodal emotion detection. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 2594–2604.

Dou Hu, Yinan Bao, Lingwei Wei, Wei Zhou, and Songlin Hu. 2023. Supervised adversarial contrastive learning for emotion recognition in conversations. *arXiv preprint arXiv:2306.01505*.

Dou Hu, Lingwei Wei, and Xiaoyong Huai. 2021a. Dialoguecrn: Contextual reasoning networks for emotion recognition in conversations. *arXiv preprint arXiv:2106.01978*.

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

Jen-tse Huang, Man Ho Lam, Eric John Li, Shujie Ren, Wenxuan Wang, Wenxiang Jiao, Zhaopeng Tu, and Michael Lyu. 2024. Apathetic or empathetic? evaluating llms' emotional alignments with humans. In *Proceedings of the Thirty-Eighth Annual Conference on Neural Information Processing Systems (NeurIPS)*.

Wenxiang Jiao, Michael Lyu, and Irwin King. 2020. Real-time emotion recognition via attention gated hierarchical memory network. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8002–8009.

Przemysław Kazienko, Julita Bielaniewicz, Marcin Gruza, Kamil Kanclerz, Konrad Karanowski, Piotr Miłkowski, and Jan Kocoń. 2023. Human-centered neural reasoning for subjective content processing: Hate speech, emotions, and humor. *Information Fusion*, 94:43–65.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kelvin JL Koa, Yunshan Ma, Ritchie Ng, and Tat-Seng Chua. 2024. Learning to generate explainable stock predictions using self-reflective large language models. In *Proceedings of the ACM on Web Conference 2024*, pages 4304–4315.

Kristin Hansen Lagattuta, Henry M Wellman, and John H Flavell. 1997. Preschoolers' understanding of the link between thinking and feeling: Cognitive cuing and emotional change. *Child development*, pages 1081–1104.

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, Fengxiao Tang, Ming Zhao, and Yusen Zhu. 2022. EmoCaps: Emotion capsule based model for conversational emotion recognition. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1610–1618, Dublin, Ireland. Association for Computational Linguistics.

Yinhan Liu. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2024. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36.

Saif Mohammad. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 174–184.

Desmond C Ong, Jamil Zaki, and Noah D Goodman. 2019. Computational models of emotion inference in theory of mind: A review and roadmap. *Topics in cognitive science*, 11(2):338–357.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Debjit Paul, Mete Ismayilzada, Maxime Peyrard, Beatriz Borges, Antoine Bosselut, Robert West, and Boi Faltings. 2023. Refiner: Reasoning feedback on intermediate representations. *arXiv preprint arXiv:2304.01904*.

Boris T Polyak and Anatoli B Juditsky. 1992. Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization*, 30(4):838–855.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386*.

Neal J Roese. 1997. Counterfactual thinking. *Psychological bulletin*, 121(1):133.

Ira J Roseman and Craig A Smith. 2001. Appraisal theory. *Appraisal processes in emotion: Theory, methods, research*, pages 3–19.

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

Sahand Sabour, Chujie Zheng, and Minlie Huang. 2022. Cem: Commonsense-aware empathetic response generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11229–11237.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

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.

Hind Taud and Jean-Franccois Mas. 2018. Multilayer perceptron (mlp). *Geomatic approaches for modeling land change scenarios*, pages 451–455.

DeepPavlov Team. 2024. Deeppavlov at semeval-2024 task 3: Multimodal large language models. In *Proceedings of the 18th International Workshop on Semantic Evaluation*, Online. Association for Computational Linguistics.

Luong Trung, Xinbo Zhang, Zhanming Jie, Peng Sun, Xiaoran Jin, and Hang Li. 2024. Reft: Reasoning with reinforced fine-tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7601–7614.

Alfredo Vellido. 2020. The importance of interpretability and visualization in machine learning for applications in medicine and health care. *Neural computing and applications*, 32(24):18069–18083.

Lisa Watson and Mark T Spence. 2007. Causes and consequences of emotions on consumer behaviour: A review and integrative cognitive appraisal theory. *European Journal of Marketing*, 41(5/6):487–511.

Sean Welleck, Ximing Lu, Peter West, Faeze Brahman, Tianxiao Shen, Daniel Khashabi, and Yejin Choi. 2022. Generating sequences by learning to self-correct. *arXiv preprint arXiv:2211.00053*.

Joshua D Wondra and Phoebe C Ellsworth. 2015. An appraisal theory of empathy and other vicarious emotional experiences. *Psychological review*, 122(3):411.

Kevin Yang, Yuandong Tian, Nanyun Peng, and Dan Klein. 2022. Re3: Generating longer stories with recursive reprompting and revision. *arXiv preprint arXiv:2210.06774*.

Weixiang Zhao, Yanyan Zhao, and Xin Lu. 2022. Cauain: Causal aware interaction network for emotion recognition in conversations. In *IJCAI*, pages 4524–4530.

A Dataset

The Third-Person Appraisal Agent was evaluated on the IEMOCAP benchmark dataset (Busso et al., 2008), which comprises conversational utterances paired with gold emotion labels. To further demonstrate the generalization capability of our framework, we evaluated it on the DailyDialog and WASSA2023 test datasets without fine-tuning. DailyDialog contains dialog-level text with previously unseen emotion labels, while WASSA2023 consists of essay-level text, requiring the agent to assess varying levels of empathy and personal distress.

IEMOCAP (Busso et al., 2008) comprises dyadic conversations between ten speakers, with the training set derived from the first eight participants. Each utterance is annotated with an emotion category.

DailyDialog (Li et al., 2017) covers various everyday topics, mirroring natural human conversation. Each utterance is annotated with an emotion category.

WASSA2023 (Hasan et al., 2023) consists of essays in which participants express their emotional reactions to news articles depicting harm to individuals, groups, or nature. Each essay is annotated with two distinct ratings based on Batson's model of empathic concern ("feeling for someone") and personal distress ("suffering with someone") (Batson et al., 1987). Empathic concern reflects the ability to understand and share another's emotions, while personal distress signifies the discomfort or anxiety experienced in reaction to another's suffering. Both ratings are measured on a 7-point scale, with 1 indicating the lowest level and 7 the highest.

B Analysis of Secondary Appraisal Phase

To demonstrate the effectiveness of the counterfactual reasoning strategy, we conduct a comparative experiment against the Reflexion-based method (Shinn et al., 2024; Koa et al., 2024). We select 500 utterances from the IEMOCAP training dataset and apply both strategies during secondary appraisal.

In Figure 2, we show the percentage change in correct predictions after each reflective iteration, using the no self-reflection (without the secondary appraisal phase) baseline as a reference. We observe that Reflexion yields moderate improvements, whereas counterfactual reasoning leads to a nearly 14.2% increase after the third iteration. This suggests that counterfactual reasoning outperforms Reflexion in enhancing correct predictions of emotions during secondary appraisal phase. One possible explanation is that Reflexion only allows the model to reflect on errors without providing specific guidance for adjustments, thus offering limited improvement in emotional reasoning.

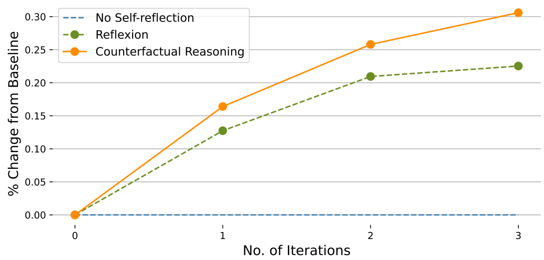

Figure 2: Percentage change in correct samples during the secondary appraisal phase, relative to the baseline values from the phase without secondary appraisal.

C Full Prompts and Their Responses

```
AppraisalInstruction_Prompt = """
Analyze the given utterance within its dialogue
    context. Provide a concise appraisal and
    predict an emotion label in the following
    format:

Situation: [Brief context description]
Speaker's perspective: [Speaker's goals or
    intentions]
Impact: [The impact of the utterance on the
    conversation]

Keep each section to 1-2 sentences. Base your
    analysis solely on the provided dialogue.
Dialogue context: {dialogue}
Utterance to analyze: {utterance}

Response Format:
Emotion Label: [choose one from: happy, sad,
    neutral, angry, excited, frustrated]
Explanation: [Brief appraisal explaining the
    chosen emotion label]

Response:
"""


Here is an example of applying
    AppraisalInstruction Prompt template.

Dialogue context: "M: Let me see them. F: Is
    there a problem? M: Who told you to get in
```

```
this line? M: Okay. But I didn't tell you
    to get in this line if you are filling out
    this particular form. F: Well what's the
    problem? Let me change it."

Utterance to analyze: "F: Well what's the
    problem? Let me change it."

prediction: neutral
appraisal: The speaker's response suggests a
    cooperative and problem-solving attitude,
    indicating a neutral emotional tone. She is
    focused on resolving the issue and
    continuing her task without delay, showing
    no apparent frustration or annoyance. Her
    offer to change lines demonstrates a
    neutral, solution-focused approach. The
    speaker's tone is matter-of-fact and
    practical, indicating a neutral emotional
    state. Her response does not convey any
    strong emotions or reactions to the
    situation. Her willingness to adapt and
    continue her task indicates a neutral,
    cooperative stance. The speaker's goal is
    to resolve the issue and proceed, without
    expressing any strong emotions. Her
    response is direct and solution-focused,
    indicating a neutral emotional tone.
```

```
CounterfactualReasoning_PROMPT = """You made
    wrong prediction, please perform a
    counterfactual analysis for the target
    utterance to refine your understanding of
    the speaker's emotional state. Follow these
    steps to guide your thinking:

1. Reflect on why predictions in
    {previous_label} mismatches between the
    predictions and the speaker's intentions or
    desires based on the target utterance.
2. Imagine an alternative emotion that better
    aligns with the speaker's intentions and
    desires based on the dialog.

Keep your analysis concise and structured. Use
    this counterfactual analysis to propose a
    more accurate emotion label that fits the
    given context.

Dialogue context: {dialog}
Utterance to analyze: {utterance}

Response Format:
Emotion Label: [choose one from: happy, sad,
    neutral, angry, excited, frustrated]
Appraisal: [Your reasoning in 2-3 short
    sentences]

Your Response:
"""

Here is an example of applying the
    Counterfactual Reasoning Prompt template.

Dialogue context: M: I don't understand why
    this is so complicated for people when they
    get here. It's just a simple form. I just
    need an ID. F: How long have you been
    working here? F: Clearly. You know, do you
    have like a supervisor or something? M:
    Yeah. Do you want to see my supervisor?
    Huh? Yeah. Do you want to see my
    supervisor? Fine. I'll be right back. F:
    That would - I would appreciate that. Yeah."

Utterance to analyze: "F: That would - I would
    appreciate that. Yeah."

###Previous trial:
previous label: neutral
previous appraisal: The female speaker uses
    polite language, such as \"I would
    appreciate that\" and \"Yeah\", which
    suggests a neutral tone, indicating she is
    seeking a resolution without expressing
    strong emotions. Her goal is to escalate
    the situation to a supervisor, implying a
    focus on finding a solution rather than
    expressing negative emotions. The polite
    tone and language used maintain a level of
    professionalism and decorum, further
    indicating a neutral emotion. The speaker's
    words and tone suggest she is trying to
    navigate the situation effectively rather
    than expressing a strong emotional state.
    Her response shows a neutral,
    matter-of-fact approach, which is in line
    with her goal of finding a solution.
    Overall, her tone and language suggest she
    is neutral and focused on finding a
    resolution.


###After counterfactual reasoning:

prediction: frustrated
adjusted appraisal:
If the speaker were feeling frustrated, it
    would suggest that she's not just seeking a
    resolution, but is also experiencing a
    sense of exasperation or annoyance with the
    situation. This could imply that she's been
    dealing with the issue for a while and is
    getting tired of the bureaucratic process.
    The frustrated tone would also indicate
    that she's not just neutral, but is
    actively seeking a solution to alleviate
    her frustration. This would change the
    dynamics of the interaction, as she might
    be more direct or forceful in her approach
    to resolve the issue.
```

```
Evaluation_PROMPT = f'''
Given the range of the class of {emotion} in
    the Circumplex Model of Affect, do the
    valence score of {valence} and the arousal
    score of {arousal} together fit within this
    range?

Answer only 'yes' or 'no'.
'''
```

```
causal_prompt = """
You are an expert in emotion classification and
    emotion cause recognition. The following is
    a conversation that involves several
    speakers. Analyze each utterance within its
    context and identify the potential cause of
    the emotion expressed in the utterance
    before predicting the emotion label.

Dialogue context: {dialogue}
Utterance to analyze: {utterance}

Response Format:
Emotion Label: [choose one from: happy, sad,
    neutral, angry, excited, frustrated]
Explanation: [Chosen emotion label based on the
    identified cause of the emotion]

Response:
"""
```

D Evaluation of Appraisal Quality

The metrics below assess the quality of emotional reasoning by evaluating the model's generated appraisals. The following descriptions detail the metrics, curated with the assistance of ChatGPT. Given the novelty of this field, research on evaluating emotional appraisals is limited.

- Sentiment Awareness

Definition: Measures the model's ability to recognize and accurately interpret the emotional tone and sentiment in communication, reflecting the speaker's feelings and attitudes.

Evaluation Criteria:

Does the appraisal effectively identify and differentiate between various emotional tones?

Does the appraisal consider the intensity of the expressed emotions?

- Contextual Understanding

Definition: Assesses the model's capacity to comprehend and integrate contextual cues when interpreting emotions.

Evaluation Criteria: Does the appraisal consider contextual cues that influence emotions?

- Sensitivity to Emotional Causes

Definition: Evaluate the model's ability to identify and understand the underlying causes of expressed emotions.

Evaluation Criteria:

Does the appraisal accurately identify and articulate the reasons or events that led to the expressed emotions?

- Emotional Dynamics Responsiveness

Definition: Assesses the model's capability to detect and respond to changes in emotional states over time.

Evaluation Criteria:

Does the appraisal effectively track and reflect changes in emotions throughout the conversation?

- Motivational Understanding

Definition: Measures the model's ability to recognize motivations of individuals behind their emotional expressions.

Evaluation Criteria:

Does the appraisal identify the speaker's motivations or goals behind their emotional state?

Does the appraisal reflect an understanding of how the speaker's emotional expressions relate to their desires or anticipated outcomes?

- Clarity and Coherence Assessment

Definition: Assess the clarity and coherence of the generated appraisals.

Evaluation Criteria:

Is the appraisal clear and easy to understand?

Does the interpretation flow coherently, linking emotional insights to contextual information?

E Appraisal Generation Comparison

Example 1:

Dialogue context: "M: I told you it was line two A. F: Right, well, last time they told me, it was this line, and obviously, it's not true. M: I don't know who told you that but they were wrong. I'm sorry. F: Do you have a manager I could speak with? F: Are you being straight with me?"

Utterance: "F: Are you being straight with me?" True label: frustrated

Third-person appraisal generator LLM: The speaker (F) is questioning the honesty of the other speaker (M), indicating frustration due to a previous misinformation about the topic (line two A) and the need for clarification. The question "Are you being straight with me?" suggests a desire for truthfulness and transparency, which may not have been present in the previous conversation. Predicted emotion label: frustrated

Original LLM: The speaker (F) is asking a direct and straightforward question, "Are you being straight with me?" without any apparent emotional tone or inflection. The language used is simple and lacks any emotional connotation, suggesting that the speaker is seeking clarity and transparency rather than expressing frustration or concern. Predicted emotion label: neutral

Example 2:

Dialogue context: "M: What is it? F: um. F: I'm sorry. It's just a lot ahh to explain. ahh I got a call– F: I got a call today, F: I'm going to need to go overseas for a while." Utterance to analyze: "F: I'm going to need to go overseas for a while." True label: sad

Third-person appraisal generator LLM: F is announcing a change in plans that involves leaving for an extended period. This change may cause separation from loved ones or disrupt current routines, which can be a source of sadness. The utterance informs M about F's plans, potentially causing a shift in the conversation's direction and possibly leading to discussions about the implications of F's departure. Predicted emotion label: sad

Original LLM: The speaker's hesitations and apologies suggest that they are frustrated with having to share the news, and the direct and simple delivery of the news implies a sense of resignation or helplessness. The speaker's tone and language convey a sense of annoyance or irritation, which is characteristic of a frustrated tone. Predicted emotion label: frustrated

Example 3:

Dialogue context: "M: You got called up. M: I thought you said this wasn't going to happen for at least a year. M: What am I going to do? F: I don't know "sigh". M: The kids-" Utterance to analyze: "M: The kids-" True label: frustrated

Third-person appraisal generator LLM: M is expressing concern about the impact of their sudden promotion on their family, particularly their children. M's utterance indicates a sense of worry and uncertainty about the future, which suggests frustration. Predicted emotion label: frustrated

Original LLM: Speaker's perspective: The speaker is introducing a relevant aspect of the situation, which is the impact on their children. The speaker mentions "The kids" as a fact, expressing a neutral emotional tone. Predicted emotion label: neutral

Example 4:

Dialogue context: "M: The kids- F: I know. I know. F: I know. Well, we'll have to figure all that out. I mean, I think the Army's going to provide childcare and help us out with that. um. M: Where do you have to go? Where are you going? F: Iraq." Utterance to analyze: "F: Iraq." True label: sad

Third-person appraisal generator LLM: The speaker (F) is expressing a location (Iraq) that is associated with military deployment, which often carries a negative connotation due to the potential dangers and separation from family. This suggests a sense of sadness or worry about the upcoming situation. Predicted emotion label: sad

Original LLM: F's utterance "Iraq" can be seen as a moment of frustration, as it represents a tangible and specific aspect of the deployment that is likely to cause worry and anxiety. The fact that F is being sent to a war zone is a difficult reality to confront, and the mention of it can be a source of frustration for both the speaker and the listener. Predicted emotion label: frustrated

Example 5:

Dialogue context: "F: We didn't see them the first time either. M: Look there, what's that? Is it? Is that a–Ah, it's a seaweed isn't it? M: Oh, this is great. I wouldn't miss this for the world. I mean think about this. I know you are not interested but think about this for a second. M: Somewhere out there, is a giant mass of silver fish all swimming in this direction. They don't know why and we don't know why. They didn't get an invitation. And it wasn't a vote- it wasn't a decision. And nobody sent them a map or anything like that. M: It's just some little instinctual time-release firecracker that went off inside them and all this one they turn around and form a line as long as the California coast and just started swimming." Utterance to analyze: "M: It's just some little instinctual time-release firecracker that went off inside them and all this one they turn around and form a line as long as the California coast and just started swimming." True label: excited

Third-person appraisal generator LLM: The speaker is describing a hypothetical scenario of a massive school of fish swimming in a coordinated manner, seemingly without a clear reason or explanation.The speaker is attempting to convey the awe-inspiring and mysterious nature of this phenomenon, likely to elicit a sense of wonder and curiosity from the listener.The utterance has a strong impact on the conversation, as it shifts the focus from a mundane observation (seaweed) to a more fantastical and intriguing idea, likely to captivate the listener's attention and imagination. The speaker's use of vivid imagery and metaphors ("time-release firecracker," "line as long as the California coast") contributes to the excitement and sense of wonder. Predicted emotion label: excited

Original LLM: The speaker's tone carries an undertone of exasperation and confusion. The phrase "I wouldn't miss this for the world" is laced with irony, hinting that they may not be particularly enthusiastic about the situation. Their struggle to grasp the reasoning behind the fish's behavior amplifies their frustration. Additionally, expressions like "some little instinctual time-release firecracker" convey a sense of resignation and helplessness, further emphasizing their lack of control over the situation. Predicted emotion label: frustrated