Appraisal Theory-Informed Emotion Prediction

Xiaowei Wang[1], Jayant Teotia[2], Rui Mao[2], Wandeep Kaur[1], Sabrina Tiun[1], Erik Cambria[2]

[1] Universiti Kebangsaan Malaysia, Malaysia

[2] Nanyang Technological University, Singapore

p143526@siswa.ukm.edu.my, {wandeep, sabrinatiun}@ukm.edu.my,

jayant002@e.ntu.edu.sg, {rui.mao, cambria}@ntu.edu.sg

Abstract

Emotion Recognition in Conversation (ERC) focuses on identifying static emotional states, overlooking the cognitive mechanisms that drive emotional transitions. This work introduces a novel emotion prediction task grounded in Appraisal Theory, which conceptualizes emotion as a cognitive evaluation of expectations and their violations. To address this task, we develop a prompt-based reasoning framework that breaks emotional dynamics into three interpretable stages, e.g., expectation inference, violation detection, and emotion-shift prediction, thereby explaining not only which emotion is expressed, but also why it emerges. To examine whether LLMs exhibit human-like affective reasoning, we design six appraisal-informed prompting tasks and evaluate eight representative LLMs across four conversational corpora. A unified two-level evaluation, which measures both emotion classification and transition dynamics, reveals that explicit expectation cues improve accuracy by up to +2.4%, whereas violation-only cues often degrade performance. Our analysis uncovers a robust appraisal pattern across models and datasets: expectation construction is the primary contributor to accurate emotion prediction, while isolated violation cues tend to induce misattribution rather than improve causal reasoning. Beyond label accuracy, transition-level evaluation shows that LLMs capture emotion-shift direction above chance but exhibit a marked stability bias, over-predicting no-change trajectories and under-detecting fine-grained shifts. These findings demonstrate both the promise and the current limits of LLMs in appraisal-driven affective reasoning, and motivate a new cognitively-grounded research direction.

Keywords: Appraisal theory, emotion prediction, large language models, cognitive science

1. Introduction

Emotions are fundamental to human cognition and behavior, consequently, modeling and understanding emotions constitute a critical aspect of artificial intelligence (Cambria et al., 2023). Emotion Recognition in Conversation (ERC) has recently emerged as a prominent direction in natural language processing research, driven by the increasing availability of large-scale conversational data (Xie and Mao, 2025; Fan et al., 2024). ERC enables the automatic inference of affective states from dialogic interactions, facilitating deeper insights into human communication dynamics. Beyond its utility for opinion mining and social media analytics, ERC holds substantial promise for affective computing applications in domains such as mental health monitoring, empathetic dialogue systems, and emotion-aware educational technologies (Mao et al., 2025).

Despite remarkable progress, current ERC research is still struggling to bridge the gap between surface-level prediction and underlying cognitive mechanisms. Early approaches based on sequential and graph neural networks (Majumder et al., 2019; Ghosal et al., 2019) have modeled contextual dependencies within dialogues, yet they largely treat emotion as a static label, lacking interpretability regarding why emotional shifts occur.

Recent causal frameworks (Li et al., 2023b; Ma et al., 2024a) attempt to infer the antecedents of emotion but fail to capture the latent cognitive variable of expectation–reality discrepancy, a key psychological driver of emotional change. Emerging appraisal-based models, such as CAPE (Liu et al., 2025) and Third-Person Appraisal Agent (Hong et al., 2025), integrated theoretical constructs from appraisal theory to enhance interpretability. However, these efforts remain fragmented and have yet to provide systematic validation of whether LLMs exhibit human-like consistency and reasoning in tracing emotional transitions.

To address these gaps, this study reconceptualizes the ERC task from the perspective of Appraisal Theory, proposing a new task, termed emotion prediction. Specifically, emotional understanding is modeled as a cognitive reasoning chain, namely expectation, violation, and emotional response. By introducing psychological variables through prompt-driven generative mechanisms, our framework enables LLMs to simulate cognitive appraisals underlying emotion shifts. Through comprehensive evaluations in multiple models and datasets, we aim to assess the ability of LLMs to understand emotions, cognitive coherence, and psychological plausibility, offering a theoretically grounded and interpretable paradigm to advance affective computing.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 11345–11358
11–16 May 2026.

This study investigates whether LLMs possess human-like emotional reasoning abilities. Through extensive experiments conducted on four multi-emotion dialog datasets, e.g., DailyDialog (Li et al., 2017), EmoryNLP (Zahiri and Choi, 2018), IEMOCAP (Busso et al., 2008), and MELD (Poria et al., 2019), we demonstrate that explicitly embedding appraisal-theoretic constructs (expectation and violation) enhances both emotion and emotion-shift prediction. Compared with the baseline condition without explicit prompts, the appraisal-informed prompts yield an average improvement of approximately 1.0% in accuracy, with the most notable gain observed on the MELD dataset (+2.4%).

The contributions of this work are twofold.

1) We introduce a novel emotion prediction task that reframes traditional ERC into a reasoning-oriented problem grounded in expectation–violation appraisal. Instead of recognizing static emotion labels, the model is required to infer how a speaker's emotional state evolves based on whether their psychological expectations are fulfilled or disrupted.

2) We systematically evaluate eight LLMs across four conversational corpora using a two-level framework measuring label accuracy and transition dynamics. Results reveal intrinsic emotional inertia and appraisal consistency biases, providing evidence of human-like affective reasoning.

2. Related Work

ERC has emerged as a central research focus for understanding emotion dynamics in human–machine interaction. Early approaches primarily treated ERC as a classification problem, leveraging recurrent and graph neural networks such as DialogueRNN and DialogueGCN to capture contextual and speaker dependencies (Majumder et al., 2019; Ghosal et al., 2019). These methods improved temporal coherence and dialogue modeling but still treated emotions as static, failing to address the challenges of similar emotions and emotion transitions. MMGCN (Hu et al., 2021) injects speaker information through speaker embeddings while capturing long-range contextual dependencies. GraphMFT (Li et al., 2023a) improves the graph construction of MMGCN, but the model misclassifies similar emotions and rushes to identify many emotions as neutral. Although GA2MIF (Li et al., 2024b) addresses the challenges of MMGCN from heterogeneous graphs, it does not solve the notorious similar emotions and emotion transfer problems in conversational emotion recognition. Although BiGMF (Lu et al., 2024) explicitly models relationships, the current ERC model still has the problem of being unable to distinguish similar emotions (such as happiness and excitement). Mao et al. (2023) found that language modeling-based methods may introduce biased predictions due to the asymmetry of emotional space segmentation. SDT (Ma et al., 2024b) proposed a transformer-based hierarchical gating fusion strategy, but it still has difficulties in distinguishing similar emotions, detecting emotions with unbalanced data, and emotion transfer. Subsequent studies began to focus on modeling emotional dynamics. Song et al. (2022) proposed EmotionFlow, which introduces an emotion propagation mechanism to simulate emotional contagion between speakers. Gao et al. (2022) first incorporated Emotion Shift Detection (ESD) as an auxiliary task within a multi-task learning framework in ESD-ERC, enabling the model to explicitly perceive emotional flow. Building on this, Wang and Mine (2023) proposed MTL-ERC-ES, which jointly learns emotion recognition, polarity, and shift detection for finer modeling of emotional transitions. Later works further extended the modality of ERC by integrating textual, acoustic, and visual cues for greater robustness. For instance, Li et al. (2024a) proposed the CFN-ESA model, which employs an emotion-shift perception module to alleviate modality discrepancies and improve recognition stability. Jian et al. (2024) advanced this line with EmoTrans, which explicitly models cross-modal emotion transitions and achieves state-of-the-art results on IEMOCAP and MELD, highlighting the importance of multimodal fusion and transition modeling.

Recent studies have further explored why emotions change. For example, MECPG jointly identifies emotional utterances and their causes to improve interpretability (Ma et al., 2024a), and EFR formulates emotion reversal as a novel reasoning task to capture triggering mechanisms (Kumar et al., 2022). Personality-aware models such as PMTL and PIRNet (Li et al., 2024c; Lian et al., 2024) have also been proposed to account for individual differences and emotional inertia across speakers. Collectively, ERC has evolved from label-based classification toward multi-dimensional psychological process modeling encompassing recognition, transition, and causation. Nevertheless, the underlying cognitive mechanisms remain underexamined (Cambria et al., 2026). Much of the literature still treats ERC as a static identification problem, even though conversational emotion is inherently dynamic and appraisal-driven: individuals form expectations, evaluate incoming utterances as confirmations or violations, and update their affective state accordingly. This motivates emotion prediction as a cognitively grounded reformulation that models how emotions emerge and shift, improving interpretability, generalization across conversational contexts, and support for emotion-aware dialogue systems.

3. Preliminary

Appraisal Theory (Scherer, 1999) states that emotional reactions are not direct outcomes of external events but are instead determined by an individual's cognitive evaluation of how those events relate to personal goals, expectations, and social norms. Emotions arise as part of a causal chain: when an individual's psychological expectations are either fulfilled or violated, an emotional shift occurs, marking a transition in affective state.

Building upon this theoretical foundation, this study develops an experimentally controlled framework. The framework employs a series of prompts that explicitly or implicitly provide psychological cues (e.g., expectations, violations) to the model. By observing the model's emotional responses across these different conditions, the study seeks to determine whether LLMs demonstrate a cognitive appraisal mechanism comparable to that observed in humans.The overall framework overview is shown in Fig. 1. The study addresses the following research questions:

RQ1: Does explicitly indicating "expectations" improve emotion prediction accuracy?

RQ2: Does indicating "violations" improve detection of emotional shifts?

RQ3: Does the combined presentation of "expectation" and "violation" information produce a synergistic improvement in performance?

RQ4: Are certain types of emotional transitions (e.g., positive→negative, goal-obstruction) more sensitive to appraisal-based mechanisms?

Based on the assumptions of Appraisal Theory, the study formulates four hypotheses:

H1: Prompts containing explicit expectation cues may improve emotion prediction accuracy compared with baseline conditions lacking such cues.

H2: Prompts containing violation cues may improve the model's ability to identify emotional shifts.

H3: Prompts combining expectation and violation cues may yield a synergistic boost.

H4: The model's consistency across multi-task outputs (measured by Cohen's κ) may serve as an indicator of whether it demonstrates a human-like appraisal-driven mechanism for emotion prediction.

4. Methodology

This section presents the experimental foundation and the datasets used in the study. Grounded in Appraisal Theory, it develops a six-stage framework of expectation, violation, and emotion shift to evaluate whether LLMs exhibit human-like patterns of emotional reasoning. To support consistent analysis across tasks, all corpora were standardized and filtered to retain instances that reflect appraisal-driven emotional dynamics.

4.1. Task Definition

Grounded in Appraisal Theory, this study conceptualizes dialogue emotion understanding as a cognitively controlled emotion prediction task rather than a static prediction task. Emotion is assumed to emerge from a speaker's appraisal of the congruence between external events and internal expectations. Accordingly, the proposed framework examines whether LLMs exhibit human-like emotional reasoning when explicit or implicit expectation cues are introduced during generation.

To empirically test these hypotheses, six interrelated subtasks (Q1–Q6) were designed to simulate the cognitive chain of expectation, violation, and emotional response. Q1 constructs the speaker's psychological expectation (E*) based on contextual cues. Q2–Q5 generate the next utterance under four cognitive conditions: baseline without cues (Q2), explicit expectation (Q3), explicit violation (Q4), and expectation-plus-violation (Q5). All models operate under identical decoding parameters to isolate cognitive effects. Finally, Q6 evaluates the emotional outputs from Q2–Q5 using independent evaluators, producing emotion labels and transition types that serve as objective measures of model sensitivity and consistency. The example utterances and example outcomes of Q1-Q6 can be viewed in Table 6 from Appendix A. These tasks form a closed cognitive-emotional modeling loop, enabling systematic verification of whether LLMs respond to expectation violations in a psychologically consistent, human-like manner.

4.2. Dataset Preparation

We developed a new dataset[1] for the evaluation of appraisal-based emotion prediction. The data were selectively sourced from four publicly available ERC datasets, e.g., MELD, IEMOCAP, EmoryNLP, and DailyDialog. These datasets differ substantially in linguistic style, discourse domain, and emotional taxonomy. Therefore, unified preprocessing and structural alignment were required before experimentation. The data were converted into a standardized CSV format and organized under a consistent indexing schema, as summarized in Table 1.

At the corpus filtering level, a two-stage selection strategy was employed to ensure both emotional dynamism and representativeness. At the dialogue level, only samples exhibiting diverse emotional distributions and evident affective variation were retained; each dialogue was required to contain at least four distinct emotion categories to avoid single-emotion bias.

[1] https://github.com/shuiguolanzi397/Emotion-Prediction-Dataset

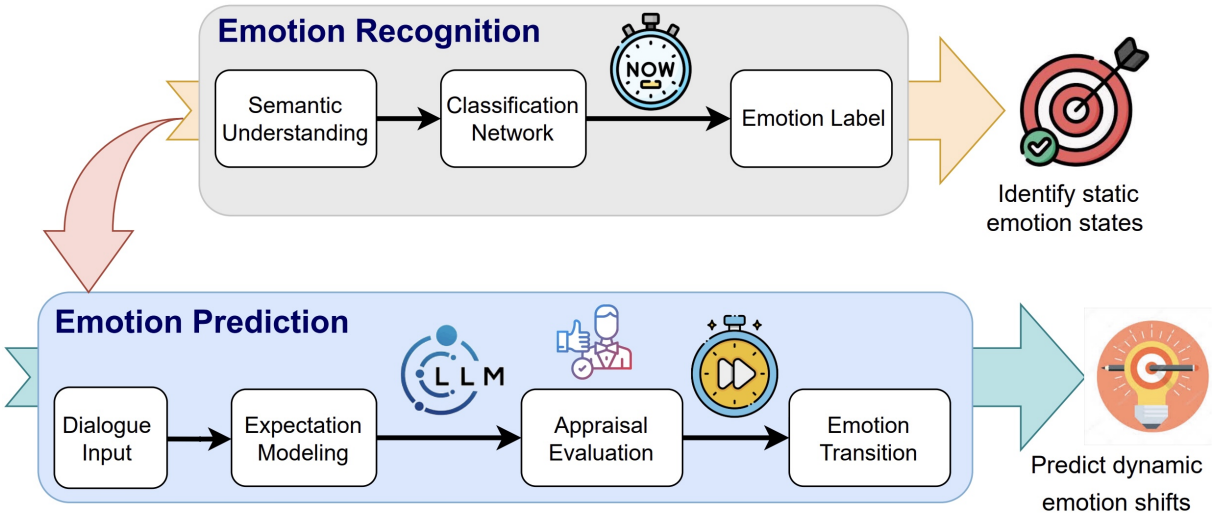

Figure 1: Framework overview: paradigm shift from emotion recognition to prediction.

ID	Description	Purpose
F1	Sample index number	Ensure sequential consistency
F2	Whether emotional shift occurs (1 = change)	Select valid samples
F3	Whether cognitively driven (1 = conforms to Appraisal Theory)	Filter non-appraisal samples
F4	Current speaker's utterance	Model input
F5	Speaker identity	Control variable (emotion inertia)
F6	Current emotion label (ground truth)	Evaluation reference
F7	Previous emotion label	Build emotion transition relation
F8	Dialogue identifier	Context alignment
F9	Utterance identifier	Mapping and traceability

Table 1: Core field definitions in the evaluation dataset. **Field abbreviations:** F1 = Sr No.; F2 = Is_Transition; F3 = Is_Appraisal_Driven; F4 = Utterance; F5 = Speaker; F6 = Emotion; F7 = Compare_With_Emotion; F8 = Dialogue_ID; F9 = Utterance_ID.

At the speaker level, only speakers who experienced at least two emotion changes within a dialogue were kept, ensuring that each retained conversation contained traceable emotional trajectories. These criteria preserve the dynamic structure of affective interactions, ensuring that model learning and evaluation occur within psychologically realistic emotional contexts.

To further mitigate structural discrepancies across corpora, dataset-specific mappings and speaker normalization were performed. For instance, in IEMOCAP, overlapping utterance indices across dyads could result in erroneous emotion chains. This issue was resolved by combining Session ID, Dyad, and Dialogue ID to generate unique conversation indices, and by standardizing speaker tags to "Speaker A/B". Moreover, only improvised sessions were retained to eliminate scripted emotional patterns, ensuring that the data more accurately reflect cognitively grounded emotional reactions in natural dialogues. After preprocessing, all four datasets were merged into a structurally consistent and label-aligned corpus covering diverse domains and emotion categories.

4.3. Quality Control

To ensure the psychological consistency and theoretical interpretability of the experimental corpus, this study designed an Appraisal-Driven Selection Mechanism for emotion-shift filtering. Unlike conventional ERC corpus cleaning or simple noise removal, this mechanism aims to automatically identify cognitively driven emotional changes within dialogues,segments that conform to the expectation–violation–emotional response logic, thereby aligning the dataset closely with the psychological processes underlying human emotion generation.

The mechanism is built upon two core indicators: Is_Transition and Is_Appraisal_Driven. The former detects intra-speaker emotional variation, while the latter determines whether such variation is triggered by another speaker's utterance. For each dialogue, the system first orders all utterances by their Dialogue ID and then iteratively backtracks to locate two types of most recent valid preceding utterances:

Algorithm 1 Determination of emotion transition and appraisal-driven mechanism

for each dialogue in dataset **do**
  sort utterances by `Utterance_ID`
  for each utterance u **do**
    $s \leftarrow u.\texttt{Speaker}$; $e \leftarrow u.\texttt{Emotion}$; $id \leftarrow u.\texttt{Utterance\_ID}$
    $prev\_self \leftarrow$ FIND_LAST$(speaker = s \wedge emotion \notin INVALID \wedge utterance_id < id)$
    if $prev\_self$ exists **then**
      $Compare\_With\_Utterance\_ID \leftarrow prev\_self.\texttt{utterance\_id}$
      $Compare\_With\_Emotion \leftarrow prev\_self.\texttt{emotion}$
      $Is\_Transition \leftarrow \mathbb{I}[e \neq prev\_self.\texttt{emotion}]$
    else
      $Compare\_With\_Utterance\_ID \leftarrow -1$
      $Compare\_With\_Emotion \leftarrow$ ""
      $Is\_Transition \leftarrow 0$
    end if
    $prev\_other \leftarrow$ FIND_LAST$(speaker \neq s \wedge emotion \notin INVALID \wedge utterance_id < id)$
    if $prev\_other$ exists **then**
      $Respond\_To\_Speaker \leftarrow prev\_other.\texttt{speaker}$
      $Respond\_To\_Utterance\_ID \leftarrow prev\_other.\texttt{utterance\_id}$
    else
      $Respond\_To\_Speaker \leftarrow$ ""
      $Respond\_To\_Utterance\_ID \leftarrow -1$
    end if
    $Is\_Appraisal\_Driven \leftarrow \mathbb{I}[Is\_Transition = 1 \wedge Respond\_To\_Speaker \neq s]$
    UPDATE(u)
  end for
end for

the same speaker's latest valid turn (the Compare-With chain) and another speaker's latest valid turn (the Respond-To chain). If the current emotion differs from the speaker's previous valid emotion, the instance is labeled Is_Transition (value 1); on this basis, if the change immediately follows a valid utterance from another speaker, it is further labeled Is_Appraisal_Driven (value 1). This dual-layer judgment ensures that each detected emotional change reflects both the speaker's internal dynamics and the external cognitive appraisal process inherent in social interaction.

To prevent self-reflective fluctuations from being misclassified as externally triggered appraisals, a Single-Speaker Run Rule was introduced. When the same speaker produces multiple consecutive utterances without any intervening turns from others, emotional variations are regarded as self-driven, with Is_Appraisal_Driven set to 0.

Similarly, at the beginning of a dialogue or in cases lacking valid preceding turns, no appraisal-driven label is assigned. This rule enforces a strict "external-trigger" principle, eliminating non-interactive emotional drifts that may occur in monologic or narrative contexts. The entire decision algorithm follows a consistent group-by-dialogue, traverse-by-utterance, nearest-backtrack, dual-judgment procedure, implemented uniformly across all datasets. The corresponding pseudocode is given in Algorithm 1.

Dataset	Original (D / U)	Kept (D / U)
DailyDialog	13118 / 102979	110 / 1380
EmoryNLP	897 / 12606	700 / 10772
IEMOCAP	74 / 5020	70 / 4851
MELD	1432 / 13708	617 / 8433

Table 2: Corpus statistics before and after corpus-level filtering across four ERC datasets. D = Dialogues, U = Utterances.

5. Experiments

5.1. Implementation Details

The experiments in this study were conducted on a hybrid hardware architecture, encompassing both local GPU deployment and remote API calls to ensure comparability and reproducibility across different model types using a unified interface. The local experiments were performed on NVIDIA A100-SXM4-40GB GPU. In terms of model deployment, offline models (DeepSeek-LLM-7B-Chat, DeepSeek-V2-Lite-Chat, Qwen-7B-Chat, Qwen2.5-14B-Instruct) are loaded through the Hugging Face Transformers framework. Online models (GPT-3.5 Turbo, GPT-4o, Gemini 1.5 Pro, Gemini 2.5 Pro) are called through the official SDK interface. The system automatically manages API keys, rate limits, and exception retry mechanisms to ensure inference stability and concurrency control. All inference hyperparameters were held constant across experiments, including a temperature of 0.7, a maximum output length of 256 tokens, and a timeout threshold of 45 seconds, to maintain comparability across different prompt conditions.

5.2. Sample Screening Based on Cognitive Mechanisms

Before conducting the formal experiments, we applied a two-level screening pipeline to the four selected ERC datasets. First, following the corpus-level filtering strategy described in Section 4.2, we obtained a reduced but emotion-dynamic subset that satisfies the dialogue-level and speaker-level constraints. Table 2 reports the dataset statistics before and after this corpus filtering step.

Second, on the corpus-filtered subset,we conducted cognitive mechanism-based sample screening and statistical analysis using two binary variables: Emotional Transition (Is_Transition) and Appraisal Driven (Is_Appraisal_Driven). The results, as shown in Fig. 2, show that appraisal-driven emotion transition dominates in all datasets, indicating that most emotional changes stem from the speaker's social evaluation of others' utterances, rather than self-reflection or emotional drift.

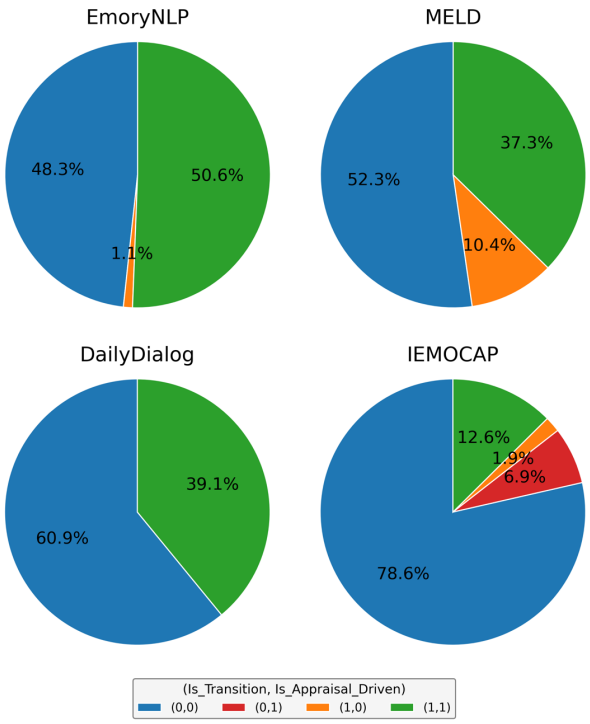

Figure 2: Sample distribution of the evaluation datasets.

Specifically, evaluation-driven emotion transitions account for approximately 39.1% (539/1380) of DailyDialog conversations, demonstrating that even in open-domain casual chat corpora, there are still numerous examples of conversations characterized by a "cognitive-emotional causal chain". EmoryNLP has the highest proportion, reaching 50.6% (5452/10772), reflecting the stronger emotional attachment between character interactions in film and television drama corpora, making cross-turn emotional resonance more likely.

In the IEMOCAP speech dialogue dataset, the original version showed a 19.5% (948/4851) rate of evaluation-driven emotion transfer. After removing ambiguous and invalid emotion labels, this rate decreased to 12.6% (611/4851), indicating a significant level of emotional uncertainty in the speech data due to tone or labeling discrepancies. The MELD rate was intermediate, with evaluation-driven emotion transfer accounting for approximately 37.3% (3148/8433), while a significant proportion of non-evaluative, self-driven emotion changes (875/8433) were also present.

Overall, these results not only provide a solid data foundation for subsequent experiments based on the "psychological expectation-appraisal outcome bias" mechanism, but also confirm that in interpersonal interactions, the evaluation of others is the primary trigger for dynamic emotion changes.

Prompt	G-1.5P	G-2.5P	GPT-3.5	GPT-4o
DailyDialog				
Q2	29.8±0.7	29.2±0.7	28.2±1.1	30.2±1.3
Q3	28.5±1.8	28.1±0.7	30.4±2.1	30.7±1.3
Q4	25.2±0.9	27.5±0.9	24.6±1.2	27.3±0.4
Q5	28.8±2.0	28.0±1.2	25.1±1.3	31.2±1.0
EmoryNLP				
Q2	27.8±1.7	27.2±1.7	27.8±1.5	28.7±1.8
Q3	28.7±1.8	28.5±2.0	29.4±1.7	29.5±1.9
Q4	26.7±1.4	27.2±2.0	26.2±1.3	27.8±1.6
Q5	28.1±1.9	28.6±2.2	26.2±1.6	29.1±1.6
IEMOCAP				
Q2	32.1±3.9	31.4±5.0	31.5±4.2	29.7±4.6
Q3	32.8±4.2	31.7±4.3	31.2±3.9	30.6±3.8
Q4	30.8±4.9	30.8±3.8	29.9±4.1	27.0±3.6
Q5	32.8±5.1	32.5±4.3	30.4±3.8	27.9±4.2
MELD				
Q2	23.3±0.5	24.3±1.0	23.3±1.1	23.8±1.0
Q3	27.1±1.0	24.9±0.9	26.6±0.6	25.6±1.3
Q4	22.6±1.5	23.5±1.2	21.2±1.3	23.3±1.2
Q5	26.5±1.2	25.7±1.3	23.5±1.4	25.2±1.4

Table 3: Accuracy (%) grouped by dataset; within each group, rows Q2–Q5 represent different prompt settings. **Abbreviations:** G-1.5P = Gemini 1.5 Pro; G-2.5P = Gemini 2.5 Pro; GPT-3.5 = GPT-3.5 Turbo; GPT-4o = GPT-4o. All values denote mean±std across evaluators.

6. Results

This study evaluates a total of eight generative models and four independent emotion evaluators across four ERC datasets under four prompt conditions (Q2–Q5). The complete accuracy matrix covering all experimental configurations is reported in Table 7 in Appendix B. Here, we focus on a representative experimental setting consisting of four generators (GPT-3.5 Turbo, GPT-4o, Gemini 1.5 Pro, Gemini 2.5 Pro) evaluated by three evaluators (GPT-4o mini, DeepSeek-V2-Lite-Chat and Qwen2.5-14B-Instruct), in order to present aggregated trends more clearly. Unless otherwise specified, all reported mean accuracy values in Sections 6.1–6.3 denote averages across these four generators, three evaluators, and the four ERC datasets.

6.1. Emotion Prediction

6.1.1. Comparison of Generative Models

Table 3 and Fig. 3 present the mean accuracy of four generative models (Gemini 1.5 Pro, Gemini 2.5 Pro, GPT-3.5 Turbo and GPT-4o), across the

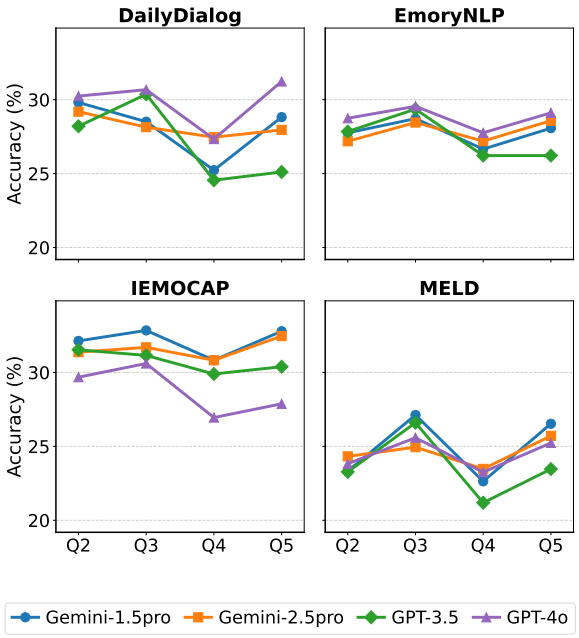

Figure 3: Mean accuracy (%) of four generators across prompts and datasets.

Prompt	G4o-m	DS-V2L	QW-14B	Mean
DailyDialog				
Q2	29.7±1.9	29.8±0.7	28.6±0.4	29.4±0.6
Q3	30.3±2.4	29.5±2.0	28.5±1.2	29.4±0.7
Q4	26.8±1.8	26.2±1.3	25.5±1.8	26.1±0.5
Q5	28.7±3.0	29.0±2.5	27.1±2.8	28.3±0.9
EmoryNLP				
Q2	30.0±0.8	27.1±0.2	26.6±1.0	27.9±1.5
Q3	31.3±0.6	28.5±0.3	27.3±1.0	29.0±1.6
Q4	28.8±1.0	26.6±0.4	25.4±0.9	27.0±1.4
Q5	30.1±1.5	27.7±1.0	26.2±1.4	28.0±1.6
IEMOCAP				
Q2	36.1±0.9	26.4±1.7	31.1±1.4	31.2±4.0
Q3	36.0±1.0	27.1±0.8	31.6±1.7	31.6±3.6
Q4	34.6±2.1	26.1±1.0	28.2±3.0	29.6±3.6
Q5	35.4±1.9	26.1±1.0	31.1±4.1	30.9±3.8
MELD				
Q2	24.1±0.8	24.4±0.5	22.5±0.5	23.7±0.8
Q3	26.8±1.3	26.5±0.6	24.9±1.2	26.1±0.8
Q4	22.7±1.0	24.0±1.0	21.2±1.1	22.6±1.1
Q5	26.2±1.3	26.0±1.2	23.5±1.4	25.2±1.2

Table 4: Accuracy (%) across datasets under four prompt conditions. **Abbreviations:** G4o-m = GPT-4o mini; DS-V2L = DeepSeek-V2-Lite-Chat; QW-14B = Qwen2.5-14B-Instruct. Values denote mean±standard deviation across evaluators.

four ERC datasets, averaged over three evaluators. Overall, the models exhibit broadly consistent performance trends under different prompting conditions. The Q3 (explicit expectation) condition often yields the most substantial improvement, whereas Q4 (violation-only) produces the most unstable results, sometimes falling below the baseline Q2 (no prompt). The Q5 (expectation + violation) condition achieves marginal gains in some corpora (e.g., DailyDialog, EmoryNLP) but does not consistently surpass Q3, suggesting that the combination of expectation and violation cues does not form a consistent synergistic effect.

Across models, GPT-4o achieves the highest mean accuracy on DailyDialog and EmoryNLP (30–32%) and remains competitive on IEMOCAP and MELD, suggesting strong cross-corpus robustness. GPT-3.5 Turbo and Gemini 2.5 Pro perform comparably (28–30%), while Gemini 1.5 Pro shows more variable performance across datasets and prompts, rather than a consistent lag. These results indicate that model scale and semantic reasoning capacity may be important for understanding expectation cues. Larger language models are better at inferring latent psychological expectations and emotional deviations from context, thereby exhibiting more human-like appraisal-driven reasoning.

Dataset-specific variations further modulate model performance. IEMOCAP achieves the highest accuracy (31–32%) owing to its high retention rate (96.6%), consistent annotations, and rich affective cues (e.g., tone, laughter, sighs). EmoryNLP ranks second (28–29%) due to dense interpersonal dependencies in multi-character scripts. In contrast, DailyDialog yields lower accuracy (27–29%) because of its limited retained size (1.34%) and short contexts, while MELD performs worst (23–26%) due to overlapping speech and ambiguous emotion attributions. These results underscore that model performance is governed not only by reasoning capability but also by corpus structure and emotional coherence: high-consistency, low-noise dialogues enable more stable and interpretable expectation-violation reasoning, whereas fragmented or low-density emotional contexts weaken the model's predictive capacity.

6.1.2. Comparison across Evaluators

Table 4 shows the mean accuracy (± std) across the four evaluation datasets and four prompt types (Q2–Q5), averaged by evaluators. Fig. 4 visualizes these trends with 95% confidence intervals. Overall, all evaluators exhibit a consistent pattern: the explicit-expectation prompt (Q3) clearly surpasses the baseline (Q2), confirming H1 that expectation cues enhance emotion prediction. Q4 (violation-only) consistently underperforms Q2 by 2–3%, indicating that violation cues alone mislead inference.

Dataset	Q2	Q3	Q4	Q5
DD	29.4±1.2	29.4±1.9	26.1±1.5	28.3±2.6
EN	27.9±1.6	29.0±1.8	27.0±1.6	28.0±2.0
IE	31.2±4.2	31.6±3.9	29.6±4.2	30.9±4.5
ME	23.7±1.0	26.1±1.3	22.6±1.5	25.2±1.7

Table 5: Mean accuracy (%) of four prompt types (Q2–Q5) averaged across all generators and evaluators. **Dataset abbreviations:** DD = DailyDialog; EN = EmoryNLP; IE = IEMOCAP; ME = MELD.

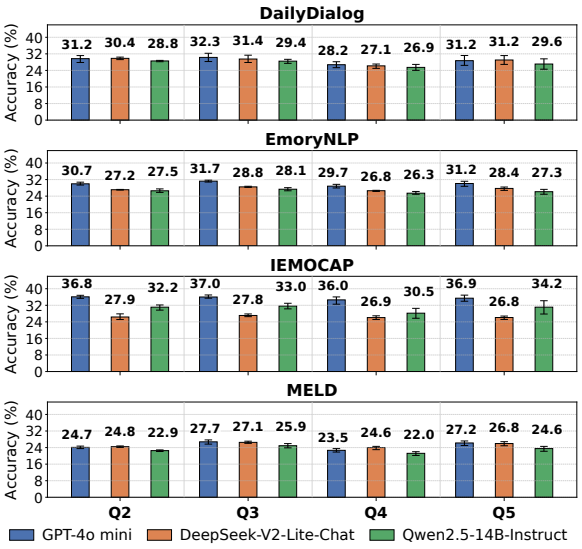

Figure 4: Mean Accuracy(%) of Q2-Q5 across three evaluators(95% CI)

Q5 (expectation + violation) achieves marginal gains over Q4 and fails to exceed Q3, offering no synergy ($\Delta H3 \approx 0$). Dataset properties further modulate these effects: IEMOCAP's low noise yields the smallest evaluator variance, whereas MELD's multi-speaker noise increases dispersion. Across all datasets, GPT-4o mini maintains the highest average accuracy but shares nearly identical relative trends with the other two evaluators, underscoring the robustness of the findings. Thus, H1 is validated, H2 and H3 are not, showing that expectation information, rather than violation cues, drives precise emotion understanding.

6.2. Prompt-level analysis

Table 5 and Fig. 5 summarize the average accuracies of four prompt types (Q2–Q5) across the four ERC datasets, averaged over four generators and three evaluators. The results reveal consistent patterns supporting the Appraisal-Theory-based hypotheses H1–H3. Explicit expectation prompts (Q3) consistently outperform the baseline (Q2) with

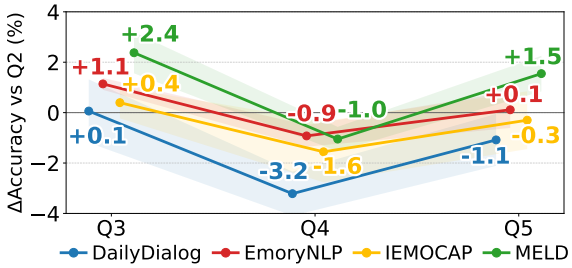

Figure 5: Relative accuracy change (Q3–Q5 vs Q2).

accuracy gains of +0.1–2.4%, confirming H1 that expectation cues enhance emotional reasoning. In contrast, violation-only prompts (Q4) lead to performance drops of 1–3%, indicating that isolated violation cues induce misattribution and weaken causal inference (disconfirming H2). When both cues co-occur (Q5), minor recoveries appear (e.g., MELD +1.5%), yet overall performance remains comparable to Q3 ($\Delta H3 \approx 0$), providing no evidence for synergy.

Dataset-level analysis shows IEMOCAP and EmoryNLP achieve the highest accuracies (30.8% / 28.0%), consistent with their higher label consistency and contextual completeness. By contrast, DailyDialog's sparse samples and MELD's multi-speaker noise limit model stability, making explicit expectation cues more crucial. Overall, the results validate that expectation construction, rather than violation detection, is the key driver of human-like emotional understanding in LLMs.

6.3. Emotional Transition Analysis

This section analyzes how explicit expectation prompts (Q3) influence emotion transition prediction from a dynamic perspective. Emotion transitions are categorized as stable (no change), upward (negative/neutral → positive), and downward (positive/neutral → negative), enabling cross-corpus comparison of emotional flow and polarity bias. As shown in Fig. 6, on EmoryNLP, GPT-4o exhibits a clear "stability bias". Both the true and predicted transition matrices highlight strong diagonal patterns dominated by Joyful→Joyful and Neutral→Neutral transitions, while the predicted results further amplify these stable routes (e.g., Neutral→Neutral rising to 492 and Joyful→Joyful to 323), indicating the model's excessive reliance on emotional inertia and limited sensitivity to shift signals. Additionally, positive emotions are frequently neutralized in prediction, as reflected by the substantial increase in Joyful→Neutral transitions (200), which suggests weak prediction of both upward and downward shifts.

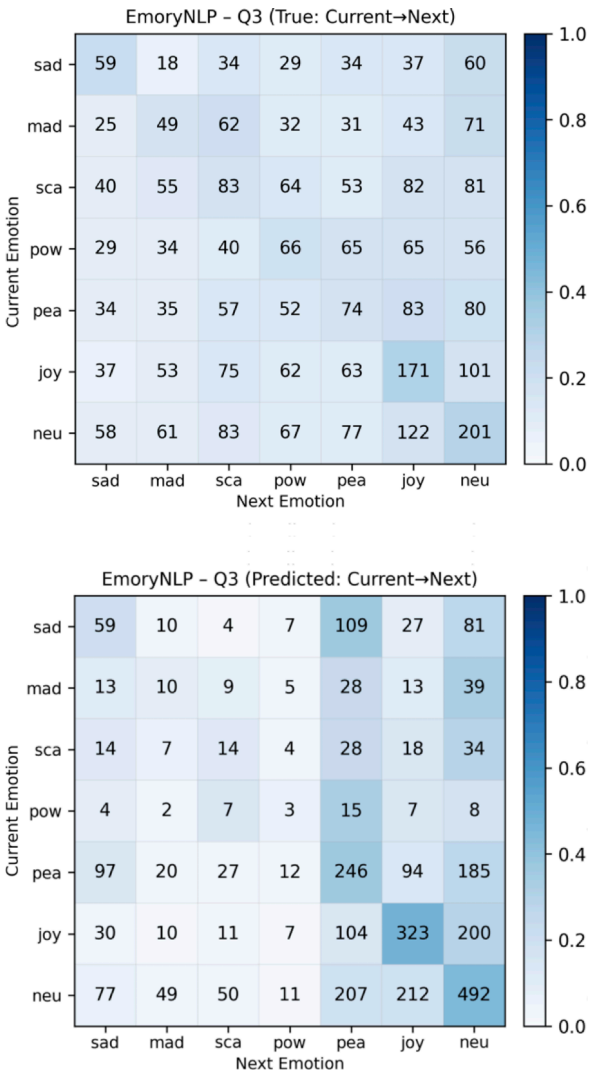

Figure 6: Q3-based transition heatmaps (colors are row-normalized proportions, numbers indicate raw counts). The labels are Sad, Mad, Scared, Powerful, Peaceful, Joyful, Neutral.

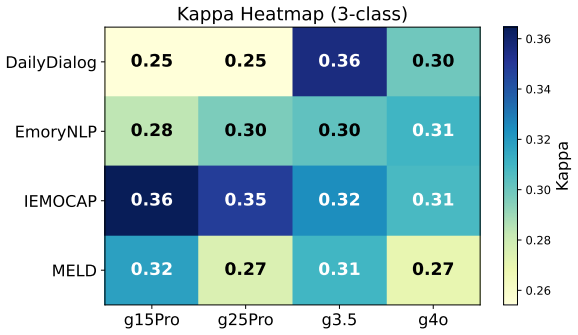

Figure 7: 3-class Cohen's κ heatmap.

Negative emotions also show systematic internal confusion, where true high-arousal transitions (e.g., among Sad, Scared and Mad) are largely diminished. Although explicit expectation prompting in Q3 improves the model's detection of "no-change" cases, fine-grained emotional dynamics in narrative contexts remain challenging to capture. In Fig. 7, all Q3 directional predictions achieve statistically significant Cohen's κ ($p < 0.001$), confirming that LLMs infer emotion-shift directions beyond chance. κ values range from 0.25 to 0.36 across datasets and generators, indicating small-to-moderate agreement beyond random polarity shifts. Agreement is consistently higher on IEMOCAP and EmoryNLP than on DailyDialog and MELD, suggesting dataset-dependent sensitivity in directional inference. Together with the transition-group recall analysis, these results provide evidence supporting H4. Detailed statistics for all dataset–generator combinations are shown in Appendix C.

7. Conclusion

This study introduces an appraisal-theory-based emotion prediction task that moves ERC beyond static emotion labels toward interpretable emotion-transition reasoning. Using a prompt-based framework with expectation inference, violation detection, and shift prediction, we assess both emotion prediction accuracy and appraisal-like reasoning in LLMs. Results consistently show that explicit expectation cues improve performance, while violation-only cues often hurt it, indicating that expectation construction is a stronger basis for affective reasoning. Combining expectation and violation cues yields no stable synergy, suggesting that more complex prompts do not always improve reasoning. These patterns are consistent across generators and evaluators, demonstrating the robustness of the framework. Transition-level analysis further shows that LLMs capture emotion-shift direction above chance, but remain biased toward emotional stability, over-predicting no-change cases and missing fine-grained shifts, especially in noisy or ambiguous multi-speaker dialogues. This reveals a clear gap between higher classification accuracy and true dynamic emotion understanding.

8. Acknowledgment

Xiaowei Wang is supported by the Putian Science and Technology Program Project (Grant No. 2025SZ3001PTXY10), funded by the Putian Municipal Science and Technology Bureau, China. Rui Mao and Erik Cambria are supported by Alibaba Group and NTU Singapore through Alibaba-NTU Global e-Sustainability CorpLab (ANGEL).

9. Bibliographical References

Erik Cambria, Rui Mao, Melvin Chen, Zhaoxia Wang, and Seng-Beng Ho. 2023. Seven pillars for the future of artificial intelligence. *IEEE Intelligent Systems*, 38(6):62–69.

Erik Cambria, Rui Mao, Amir Hussain, Keith Oatley, and Geoffrey Hinton. 2026. Artificial intelligence as the fourth decentering revolution: From cosmic, biological, and psychological displacement to cognitive decentering. *Cognitive Computation*, 18(20):1–13.

Chunxiao Fan, Jie Lin, Rui Mao, and Erik Cambria. 2024. Fusing pairwise modalities for emotion recognition in conversations. *Information Fusion*, 106:102306.

Qingqing Gao, Biwei Cao, Xin Guan, Tianyun Gu, Xing Bao, Junyan Wu, Bo Liu, and Jiuxin Cao. 2022. Emotion recognition in conversations with emotion shift detection based on multi-task learning. *Knowledge-Based Systems*, 248:108861.

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*, pages 154–164.

Simin Hong, Jun Sun, and Hongyang Chen. 2025. Third-person appraisal agent: Simulating human emotional reasoning in text with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 23684–23701.

Jingwen Hu, Yuchen Liu, Jinming Zhao, and Qin Jin. 2021. MMGCN: Multimodal fusion via deep graph convolution network for emotion recognition in conversation. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5666–5675.

Zhongquan Jian, Ante Wang, Jinsong Su, Junfeng Yao, Meihong Wang, and Qingqiang Wu. 2024. EmoTrans: Emotional transition-based model for emotion recognition in conversation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, pages 5723–5733.

Shivani Kumar, Anubhav Shrimal, Md Shad Akhtar, and Tanmoy Chakraborty. 2022. Discovering emotion and reasoning its flip in multi-party conversations using masked memory network and transformer. *Knowledge-Based Systems*, 240:108112.

Jiang Li, Xiaoping Wang, Yingjian Liu, and Zhigang Zeng. 2024a. CFN-ESA: A cross-modal fusion network with emotion-shift awareness for dialogue emotion recognition. *IEEE Transactions on Affective Computing*, 15(4):1919–1933.

Jiang Li, Xiaoping Wang, Guoqing Lv, and Zhigang Zeng. 2023a. GraphMFT: A graph network based multimodal fusion technique for emotion recognition in conversation. *Neurocomputing*, 550:126427.

Jiang Li, Xiaoping Wang, Guoqing Lv, and Zhigang Zeng. 2024b. GA2MIF: Graph and attention based two-stage multi-source information fusion for conversational emotion detection. *IEEE Transactions on Affective Computing*, 15(1):130–143.

Wei Li, Liuyao Zhu, Rui Mao, and Erik Cambria. 2023b. SKIER: A symbolic knowledge integrated model for conversational emotion recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(11):13121–13129.

Xiaonan Li, Dandan Song, and Yanru Zhou. 2024c. PMTL: Personality-aware multitask learning model for emotion recognition in conversation. In *2024 6th International Conference on Communications, Information System and Computer Engineering (CISCE)*, pages 428–432.

Zheng Lian, Bin Liu, and Jianhua Tao. 2024. PIRNet: Personality-enhanced iterative refinement network for emotion recognition in conversation. *IEEE Transactions on Neural Networks and Learning Systems*, 35(2):2863–2874.

June M Liu, He Cao, Renliang Sun, Rui Wang, Yu Li, and Jiaxing Zhang. 2025. CAPE: A Chinese dataset for appraisal-based emotional generation in large language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 6291–6309.

Nannan Lu, Zhiyuan Han, Min Han, and Jiansheng Qian. 2024. Bi-stream graph learning based multimodal fusion for emotion recognition in conversation. *Information Fusion*, 106:102272.

Heqing Ma, Jianfei Yu, Fanfan Wang, Hanyu Cao, and Rui Xia. 2024a. From extraction to generation: Multimodal emotion-cause pair generation in conversations. *IEEE Transactions on Affective Computing*, pages 1–12.

Hui Ma, Jian Wang, Hongfei Lin, Bo Zhang, Yijia Zhang, and Bo Xu. 2024b. A Transformer-based model with self-distillation for multimodal emotion recognition in conversations. *IEEE Transactions on Multimedia*, pages 1–13.

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.

Rui Mao, Mengshi Ge, Sooji Han, Wei Li, Kai He, Luyao Zhu, and Erik Cambria. 2025. A survey on pragmatic processing techniques. *Information Fusion*, 114:102712.

Rui Mao, Qian Liu, Kai He, Wei Li, and Erik Cambria. 2023. The biases of pre-trained language models: An empirical study on prompt-based sentiment analysis and emotion detection. *IEEE Transactions on Affective Computing*, 14(3):1743–1753.

Klaus R. Scherer. 1999. Appraisal theory. In *Handbook of Cognition and Emotion*, pages 637–663. John Wiley & Sons Ltd., Chichester, UK.

Xiaohui Song, Liangjun Zang, Rong Zhang, Songlin Hu, and Longtao Huang. 2022. EmotionFlow: Capture the dialogue level emotion transitions. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8542–8546.

Juntao Wang and Tsunenori Mine. 2023. Multi-task learning for emotion recognition in conversation with emotion shift. In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, pages 257–266.

Yunhe Xie and Rui Mao. 2025. PGIF: A personality-guided iterative feedback graph network for multimodal conversational emotion recognition. *IEEE Transactions on Computational Social Systems*, 12(5):3583–3595.

10. Language Resource References

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.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing*, pages 986–995. Asian Federation of Natural Language Processing.

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.

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.

A. Task Definition Example

To examine how cognitive framing influences emotional response generation, we instantiate the proposed framework within a controlled dialogue setting derived from the DailyDialog dataset. By holding decoding parameters constant across Q2–Q5, we ensure that any observed variation in emotional outcomes arises from differences in cognitive guidance rather than stochastic decoding effects. Emotion labeling in Q6 is performed independently by a separate evaluator model, enabling objective assessment of emotional consistency across conditions. The response utterances in Q2–Q5 were generated using the GPT-4o model under identical decoding parameters, while emotional evaluation in Q6 was conducted using the Gemini 2.5 Flash model (see Table 6). The dialogue context underlying this illustrative example is provided below.

B1: the food station was as crowded as usual. I waited for 20 minutes for my sandwich , and when I finally got it the chicken was underdone .
A1: I see. You simply can't blame the server for this.
B2: yes, I know. Fast food is so popular in this district.
A2: that's true.
B3: anyway, how was your lunch, Catherine?
A3: it was good. I brought lunch from home today.

Within this example, Q1 reconstructs Speaker B's expectation at the B3 time point, conditioned on the observed B4 utterance. This step operationalizes the latent cognitive frame associated with the realized outcome and is conducted offline for analytical purposes.

Task	Theoretical Motivation (Why)	Prompting & Operational Design (How)	Example Output (Same Dialogue Instance)
Q1	Constructs the speaker's anticipatory cognitive frame. In appraisal theory, emotions arise from how individuals interpret events relative to their internal understanding.	Infer what speaker was expecting based on dialogue context. Provide a neutral description of the likely expectation.	B expected A to have eaten lunch at a fast food restaurant, similar to those in America.
Q2	Cognitive-neutral baseline to measure spontaneous emotional tendencies without explicit expectation or violation cues.	Generate the next utterance using only dialogue context. Identical decoding parameters across Q2–Q5.	"That's a smart move. I should start doing that too to avoid the rush and the disappointment."
Q3	Tests whether explicit activation of prior expectation influences emotional response generation.	Provide the extracted expectation during generation. The response must reflect speaker's expectation.	"Oh, I see. I was expecting you might have gone to one of those fast food places like we have back in America."
Q4	Examines whether prompting implicit evaluation of outcome alignment influences emotional coherence.	Instruct the model that speaker's emotion may depend on whether interlocutor's reply matched or did not match what speaker hoped for. No explicit expectation text is provided.	"Oh, that's nice. Maybe I should start bringing my own lunch too. It might save me some time and hassle."
Q5	Represents a fully appraisal-informed condition combining anticipatory framing and outcome evaluation to guide emotion prediction.	Provide both the extracted expectation and evaluative guidance regarding match/mismatch during generation. Generate an emotionally coherent response.	"Oh, you brought your lunch? That's smart! I should start doing that too, especially after today's experience."
Q6	Provides an independent evaluation of emotional consistency across cognitively guided conditions. Within appraisal-informed prediction, emotional outcomes should reflect differences in cognitive framing.	Separate evaluator labels emotions from Q2–Q5 outputs. This step is analytical only and does not influence generation.	Predicted emotions: Q2: sadness Q3: surprise Q4: neutral Q5: happiness (Ground truth: surprise)

Table 6: Motivation, prompt and output of Q1–Q6 with illustrative examples on the same dialogue instance. Identical decoding parameters are maintained across Q2–Q5.

Q2–Q5 then generate Speaker B's next utterance (at the B4 step) after observing A3, under different cognitive guidance conditions. Importantly, during this generation stage, the model does not have access to the ground-truth B4. Finally, Q6 assigns emotion labels to the generated B4 responses from Q2–Q5 to assess emotional sensitivity and coherence. This evaluation step is purely analytical and does not influence generation. Among the emotion labels assigned by Q6 to the outputs of Q2–Q5 in the given example, only the label predicted for Q3 is consistent with the ground truth in Table 6.

B. Full Experimental Results

We report complete accuracy results across all experimental configurations in Table 7. The full matrix includes eight generative models (GPT-3.5 Turbo, GPT-4o, Gemini 1.5 Pro, Gemini 2.5 Pro, DeepSeek-LLM-7B-Chat, DeepSeek-V2-Lite-Chat, Qwen-7B-Chat and Qwen2.5-14B-Instruct), four independent evaluators (GPT-4o mini, Gemini 2.5 Flash, DeepSeek-V2-Lite-Chat and Qwen2.5-14B-Instruct), four ERC datasets (DailyDialog, EmoryNLP, IEMOCAP and MELD), and four prompt conditions (Q2–Q5). We further provide a structured statistical analysis of the full matrix to quantify prompt-effect stability, model-level variation, dataset-conditioned behavior, evaluator robustness, and corpus-level difficulty ordering across configurations.

Across the complete experimental matrix reported in Table 7 (8 generators $\times$ 4 evaluators $\times$ 4 datasets $\times$ 4 prompt conditions; 128 total configurations), the observed prompt effects exhibit systematic structure rather than stochastic fluctuation. Globally, Q3 outperforms Q2 in 98 out of 128 configurations (76.6%), while Q4 underperforms Q2 in 102 out of 128 configurations (79.7%). In contrast, Q5 exceeds Q3 in only 26 out of 128 configurations (20.3%). These proportions indicate that expectation-guided prompting (Q3) yields a robust improvement over the baseline (Q2), violation-only prompting (Q4) consistently degrades performance,

Model	Cond.	DailyDialog				EmoryNLP				IEMOCAP				MELD			
		GPT	GEM	DS	QW	GPT	GEM	DS	QW	GPT	GEM	DS	QW	GPT	GEM	DS	QW
DS-7B	Q2	24.68%	24.86%	27.64%	24.12%	28.98%	23.92%	27.07%	25.22%	35.35%	33.55%	27.17%	31.26%	22.78%	22.05%	**24.24%**	21.12%
	Q3	**29.87%**	**25.97%**	**29.31%**	**25.79%**	**29.53%**	**24.58%**	**27.11%**	**26.49%**	**36.82%**	**34.70%**	26.35%	**31.75%**	**23.09%**	**24.05%**	23.70%	**22.11%**
	Q4	27.27%	24.12%	28.20%	25.60%	28.52%	23.70%	25.48%	25.13%	34.37%	30.28%	26.19%	28.15%	21.41%	21.41%	22.62%	19.70%
	Q5	23.75%	24.49%	24.30%	22.45%	28.19%	23.90%	26.72%	24.63%	34.86%	34.53%	**28.48%**	31.42%	21.28%	21.89%	22.90%	18.90%
DS-V2L	Q2	25.60%	28.39%	28.39%	25.79%	30.14%	24.36%	26.83%	26.41%	**35.02%**	33.06%	25.70%	**30.28%**	23.67%	23.09%	24.75%	22.36%
	Q3	**28.57%**	**28.76%**	**31.35%**	**27.27%**	**31.58%**	**26.30%**	27.77%	**27.90%**	**35.02%**	**34.04%**	**27.50%**	28.97%	**24.43%**	**25.03%**	**25.57%**	**22.52%**
	Q4	26.35%	24.12%	26.72%	22.45%	30.32%	25.09%	26.52%	26.89%	32.90%	33.06%	25.53%	28.81%	21.92%	21.19%	23.25%	20.74%
	Q5	25.23%	26.35%	28.01%	24.68%	31.11%	26.01%	**28.08%**	26.85%	34.21%	31.91%	26.02%	28.97%	23.57%	22.84%	24.46%	20.49%
G-1.5P	Q2	**30.06%**	**32.28%**	30.43%	**28.94%**	29.89%	25.64%	27.08%	26.29%	36.99%	31.42%	**28.64%**	30.77%	23.36%	24.54%	23.88%	22.74%
	Q3	27.83%	31.54%	30.80%	26.90%	**30.93%**	**26.99%**	**28.25%**	**26.97%**	**37.48%**	**32.24%**	28.15%	32.90%	**28.06%**	**29.04%**	**27.37%**	**25.93%**
	Q4	24.68%	26.35%	26.35%	24.68%	28.16%	23.81%	26.83%	24.98%	36.33%	31.42%	25.37%	30.77%	22.26%	23.23%	24.42%	21.23%
	Q5	27.46%	28.76%	**31.35%**	27.64%	30.12%	26.90%	28.14%	25.95%	**37.48%**	31.59%	26.51%	**34.37%**	27.47%	27.83%	27.19%	24.93%
G-2.5P	Q2	**30.06%**	**31.17%**	**28.94%**	**28.57%**	29.18%	25.04%	26.83%	25.53%	**36.66%**	34.21%	25.53%	31.91%	25.06%	26.56%	24.90%	23.00%
	Q3	28.76%	30.80%	27.27%	28.39%	30.67%	25.31%	**28.48%**	26.21%	35.68%	**35.52%**	26.35%	33.06%	25.13%	**27.70%**	25.86%	23.86%
	Q4	28.57%	28.76%	27.09%	26.72%	29.57%	24.98%	26.78%	25.24%	35.35%	33.06%	26.84%	30.28%	23.51%	24.46%	24.75%	22.17%
	Q5	29.50%	30.98%	27.46%	26.90%	**31.24%**	**25.55%**	27.95%	**26.54%**	36.33%	34.04%	**27.00%**	**34.04%**	**26.65%**	27.35%	**26.43%**	**24.05%**
GPT-3.5	Q2	27.09%	30.61%	**29.50%**	28.01%	29.71%	25.50%	27.20%	26.61%	**35.68%**	**35.19%**	26.51%	**32.41%**	23.67%	23.98%	24.27%	21.89%
	Q3	**33.02%**	**31.17%**	28.39%	**29.68%**	**31.42%**	**25.57%**	**28.94%**	**27.71%**	35.52%	32.73%	26.84%	31.10%	**27.29%**	**27.86%**	**26.59%**	**25.92%**
	Q4	25.97%	25.97%	24.30%	23.38%	27.73%	23.46%	26.10%	24.82%	35.19%	29.95%	**27.00%**	27.50%	21.41%	22.14%	22.52%	19.63%
	Q5	25.42%	25.23%	26.35%	23.56%	27.95%	23.70%	26.32%	24.38%	34.70%	33.22%	26.19%	30.28%	24.43%	25.32%	24.36%	21.60%
GPT-4o	Q2	31.54%	31.54%	30.43%	28.76%	31.03%	25.97%	27.29%	27.90%	35.02%	32.08%	24.71%	29.30%	24.30%	24.40%	24.68%	22.49%
	Q3	31.54%	**32.28%**	**31.54%**	28.94%	**31.97%**	**27.16%**	28.21%	**28.47%**	**35.35%**	**33.88%**	**27.00%**	**29.46%**	**26.62%**	**27.41%**	**26.24%**	**23.89%**
	Q4	27.83%	27.46%	27.09%	27.09%	29.82%	25.46%	26.76%	26.69%	31.59%	31.10%	25.04%	24.22%	23.51%	23.28%	24.40%	21.86%
	Q5	**32.47%**	30.98%	30.98%	**30.24%**	31.13%	26.10%	**28.48%**	27.71%	33.22%	30.28%	24.71%	25.70%	26.37%	26.68%	25.83%	23.51%
QW-7B	Q2	**24.68%**	**25.97%**	27.83%	21.71%	28.80%	**25.29%**	26.58%	24.56%	35.35%	34.70%	**27.99%**	**32.90%**	21.32%	21.70%	24.02%	19.98%
	Q3	23.38%	25.23%	27.83%	23.56%	**30.32%**	24.65%	**27.55%**	**26.17%**	**35.52%**	**35.68%**	26.19%	32.24%	22.74%	**23.82%**	**25.10%**	**21.57%**
	Q4	21.71%	22.26%	**28.20%**	23.01%	28.47%	23.79%	26.63%	25.24%	**35.52%**	**35.68%**	26.35%	31.59%	21.06%	20.71%	24.59%	19.25%
	Q5	21.89%	24.30%	25.42%	**23.75%**	29.13%	24.43%	26.98%	25.97%	34.70%	34.53%	26.19%	32.08%	**22.90%**	23.79%	23.79%	20.97%
QW-14B	Q2	28.20%	**29.87%**	**30.43%**	**29.31%**	28.63%	24.65%	26.23%	25.26%	36.17%	**34.37%**	26.51%	30.93%	21.89%	22.62%	23.89%	21.16%
	Q3	**28.94%**	29.13%	29.87%	**29.31%**	**30.32%**	**25.48%**	**28.41%**	**25.70%**	35.19%	32.90%	27.00%	30.77%	**26.02%**	**26.94%**	**25.13%**	**23.35%**
	Q4	26.16%	24.86%	25.79%	23.56%	27.29%	22.91%	26.39%	23.29%	**36.82%**	33.39%	**27.50%**	31.10%	20.78%	20.74%	21.98%	18.46%
	Q5	26.35%	26.16%	28.20%	25.05%	27.95%	23.84%	28.14%	24.27%	35.19%	34.04%	**27.50%**	**33.22%**	22.11%	25.16%	23.35%	19.38%

Table 7: Full results (%) across datasets and prompting conditions (Q2–Q5). **Cond.** denotes prompting condition (Q2–Q5). **Formatting: Bold** indicates the best score among Q2–Q5 within each generator for the same column. **Bold-underlined** indicates the overall best score in that column across all generators. **Evaluator abbreviations:** GPT = GPT-4o mini; GEM = Gemini 2.5 Flash; DS = DeepSeek-V2-Lite-Chat; QW = Qwen2.5-14B-Instruct. **Generator abbreviations:** DS-7B = DeepSeek-LLM-7B-Chat; DS-V2L = DeepSeek-V2-Lite-Chat; G-1.5P = Gemini 1.5 Pro; G-2.5P = Gemini 2.5 Pro; GPT-3.5 = GPT-3.5 Turbo; GPT-4o = GPT-4o; QW-7B = Qwen-7B-Chat; QW-14B = Qwen2.5-14B-Instruct.

and the hypothesized synergistic effect of combining expectation and violation (Q5) is not generally supported.

Dataset-conditioned analysis reveals marked differences in stability. The improvement Q3 > Q2 holds in 31/32 configurations (96.9%) on both EmoryNLP and MELD, but drops to 17/32 (53.1%) on DailyDialog and 19/32 (59.3%) on IEMOCAP. Similarly, Q4 < Q2 remains highly consistent across datasets, with the strongest stability on MELD (30/32, 93.8%). These results suggest that the effectiveness of expectation-guided prompting is corpus-sensitive, with stronger regularity in datasets exhibiting clearer affective trajectories under the applied filtering criteria.

Generator-level aggregation further clarifies model dependence. Q3 > Q2 holds in 15/16 configurations (93.8%) for GPT-4o, 14/16 (87.5%) for both DeepSeek-V2-Lite-Chat and DeepSeek-LLM-7B-Chat, but decreases to 10/16 (62.5%) for Qwen-7B-Chat and Qwen2.5-14B-Instruct. Global mean accuracy averaged over all settings shows a narrow performance band among top-tier models (Gemini 2.5 Pro: 0.2832; Gemini 1.5 Pro: 0.2826; GPT-4o: 0.2811), followed by GPT-3.5 Turbo (0.2731) and smaller open models (0.264–0.271). However, dataset-specific rankings vary: GPT-4o leads on DailyDialog and EmoryNLP, Qwen-7B-Chat marginally leads on IEMOCAP, and Gemini 1.5 Pro leads on MELD. These variations indicate that scaling advantages are not uniformly realized across corpora.

Evaluator-level comparisons demonstrate strong ranking stability despite calibration differences. Averaged over generators and datasets, all evaluators yield the identical prompt ordering Q3 > Q2 > Q5 > Q4. Absolute accuracy levels vary, with GPT-4o mini producing the highest overall mean (0.2903), followed by Gemini 2.5 Flash (0.2759), DeepSeek-V2-Lite-Chat (0.2668), and Qwen2.5-14B-Instruct (0.2618). Mean prompt deltas are directionally consistent across evaluators (mean Q3–Q2 $\approx$ +0.00963; mean Q4–Q2 $\approx$ -0.01441; mean Q5–Q3 $\approx$ -0.01208), indicating that evaluator

Model	N	κ	CI95_low	CI95_high
DailyDialog				
Gemini 1.5 Pro	199	0.2542	0.1519	0.3558
Gemini 2.5 Pro	199	0.2544	0.1539	0.3513
GPT-3.5 Turbo	199	0.3555	0.2556	0.4578
GPT-4o	199	0.2989	0.1955	0.4005
EmoryNLP				
Gemini 1.5 Pro	3140	0.2831	0.2561	0.3099
Gemini 2.5 Pro	3042	0.2962	0.2690	0.3230
GPT-3.5 Turbo	3045	0.3024	0.2738	0.3293
GPT-4o	3043	0.3094	0.2808	0.3377
IEMOCAP				
Gemini 1.5 Pro	102	0.3646	0.2114	0.5158
Gemini 2.5 Pro	100	0.3476	0.1949	0.4946
GPT-3.5 Turbo	102	0.3237	0.1577	0.4730
GPT-4o	102	0.3075	0.1364	0.4676
MELD				
Gemini 1.5 Pro	2008	0.3172	0.2841	0.3484
Gemini 2.5 Pro	1433	0.2749	0.2332	0.3136
GPT-3.5 Turbo	1441	0.3093	0.2674	0.3500
GPT-4o	1441	0.2701	0.2307	0.3101

Table 8: Agreement (Cohen's κ) with 95% confidence intervals across datasets and models. Permutation test yields $p_{\text{perm}} = 0.00019996$ for all configurations.

Model	Down R	Stable R	Up R	IB	PB	NB
DailyDialog						
G-1.5P	10.3	1.4	6	-12.6	-6	-2.5
G-2.5P	1.7	2.7	3	-8.5	-8.5	-2
GPT-3.5	1.7	2.7	1.5	-9	-8	-4.5
GPT-4o	3.4	1.4	1.5	-9.5	-6	-2
EmoryNLP						
G-1.5P	6.2	13.6	6.9	-0.9	-0.4	0.9
G-2.5P	7.7	12.4	7.3	-1.6	0	1.4
GPT-3.5	7.6	19.2	8.5	1	-0.5	-0.6
GPT-4o	7.7	22.4	7.5	3.7	-2	-1.8
IEMOCAP						
G-1.5P	5.6	59.3	0	27.5	-17.6	-15.7
G-2.5P	0	61	0	25.5	-12.7	-16.7
GPT-3.5	0	59.3	4	31.4	-18.6	-13.7
GPT-4o	0	54.2	0	22.5	-14.7	-9.8
MELD						
G-1.5P	6.9	3.2	6.3	12	-7.1	-5.4
G-2.5P	6.3	8.1	4.7	3	-1	-2.5
GPT-3.5	5	1.6	6.1	5.6	-2	-3.5
GPT-4o	3.7	1.6	2.8	6	-3	-3.1

Table 9: Recall for downward (Down R), stable (Stable R), and upward (Up R) transitions, with imbalance (IB), positive bias (PB), and negative bias (NB). **Model abbreviations:** G-1.5P = Gemini 1.5 Pro; G2.5P = Gemini 2.5 Pro; GPT-3.5 = GPT-3.5 Turbo; GPT-4o = GPT-4o.

choice affects absolute calibration but not relative prompt effects.

Dataset difficulty ordering is highly stable. IEMOCAP ranks first (highest accuracy) in 96/128 configurations, while MELD ranks last in 112/128 configurations. EmoryNLP and DailyDialog frequently alternate between second and third positions. This pattern suggests consistent relative difficulty extremes with intermediate variability.

C. Statistical Robustness and Bias Analysis of Directional Agreement

Table 8 reports full Cohen's κ statistics for all 16 dataset–generator combinations under Q3 prompting using the GPT-4o mini emotion evaluator. Directional agreement was computed over three transition classes (Up, Stable, Down), with N denoting the number of aligned ground-truth and predicted transition pairs used for evaluation.

All configurations yield statistically significant agreement beyond chance (permutation test, $p_{\text{perm}} < 0.001$), and the 95% confidence intervals do not cross zero. κ values range from 0.25 to 0.36, indicating fair-to-moderate agreement across datasets and models. Together, these results confirm that directional consistency cannot be explained by random polarity shifts but reflects structured affective inference.

To examine whether directional agreement is driven by structural transition bias, Table 9 reports transition-group recall and bias indicators, including Inertia Bias (IB), Positivity Bias (PB), and Negativity Bias (NB). Positive bias values indicate overprediction relative to the empirical class distribution, whereas negative values indicate under-prediction.

Models exhibit measurable stability preference on certain datasets, particularly IEMOCAP (IB > 20%), indicating partial reliance on emotional inertia. However, bias patterns vary across corpora: IB is near zero or negative on EmoryNLP and DailyDialog, where non-trivial upward and downward recall remains observable. These results suggest that directional agreement cannot be reduced to simple stability dominance but reflects structured, dataset-dependent affective inference.