Exposing Weaknesses in Emotion Recognition in Conversations

Anonymous ACL submission

Abstract

Emotion Recognition in Conversations (ERC) aims to identify speakers' emotions in multi-turn dialogue. While many recent approaches rely on task-specific fine-tuning, such models may exploit dataset-specific cues. To reduce this effect, we study ERC using Large Language Models (LLMs) in a zero-shot setting, incorporating preceding conversational turns as context. Across standard ERC benchmarks, aggregate evaluation metrics mask substantial differences in per-class behavior despite strong overall performance. We observe that errors occur more frequently for utterances with short replies, interjections, negations, and sentence-type markers such as exclamations and interrogatives. These error patterns raise the question of whether they reflect model behavior or properties of the benchmark datasets themselves. To further investigate this issue, we conduct a controlled re-annotation study with four additional human annotators, treating the original dataset annotation as a fifth annotator. Strong annotator agreement is observed in only 35% of cases, with (80%) neutral utterances accounting for most high-agreement instances, indicating that emotion plausibility is a central issue in ERC evaluation. Finally, we analyze model behavior across different agreement levels and introduce an LLM-as-Judge framework that explicitly evaluates emotion plausibility, allowing multiple emotionally coherent interpretations rather than enforcing a single-label decision.

1 Introduction

Emotion Recognition in Conversations (ERC) is the task of predicting the emotion of each utterance in multi-turn dialogues (Majumder et al., 2019; Ghosal et al., 2019). ERC is considered a significant research direction due to its wide applications such as empathetic dialogue systems (Rashkin et al., 2019; Zhou et al., 2018), mental health monitoring (Cummins et al., 2019; Li et al., 2023b),

Dialogue:

Ross: So… what did they say after the interview?
Rachel: I don't know. They smiled, thanked me, and said they'd call.
Rachel: **Yeah. I guess that's something!**

Labels: *Anger, Disgust, Fear, Joy, Neutral, Sadness, Surprise.*

Original 1: Neutral
Annotator 2: Sadness
Annotator 3: Fear
Annotator 4: Joy
Annotator 5: Sadness

Interpretation:

Example of an **ambiguous conversational utterance** where emotional interpretation depends on implicit intent and pragmatic context. T

Figure 1: Example of annotation ambiguity in ERC due to multiple valid emotional interpretations.

and is crucial for machines to understand dynamic human emotions. This deeper comprehension has meaningful impacts, such as supporting mental health monitoring and improving human–machine interaction. As people increasingly interact with machines in their daily lives, it has become natural to expect these systems not only to understand the content of what is said, but also to recognize the emotions conveyed and to respond in ways that align with those emotional cues. Research in ERC has been shaped by datasets such as MELD, EmoryNLP, DailyDialog, and IEMOCAP (Poria et al., 2019; Zahiri and Choi, 2018; Li et al., 2017; Busso et al., 2008). Classical methods (e.g., DialogueRNN, DialogueGCN, COSMIC) modeled speaker states and conversational structures (Majumder et al., 2019; Ghosal et al., 2019, 2020). More recently, Large Language Models (LLMs) have been adopted, typically via fine-tuning or few-shot prompting (Lei et al., 2023; Shen et al., 2025). Fine-tuned models tend to perform well within their

training domain when provided with auxiliary information such as scene descriptions or persona details, but they struggle to generalize to datasets where such contextual signals are absent (Shen et al., 2025; Fu et al., 2024; Chen and Xiao, 2024).

Beyond dataset-level differences, ERC evaluation commonly assumes that each utterance can be assigned a single, unambiguous gold emotion label. However, emotional expressions in conversation are frequently shaped by contextual interpretation, intention, and linguistic cues, making strict single-label annotation insufficient in many cases. This raises an *important question*: when models appear to fail on ERC benchmarks in terms of recognition performance, do these errors necessarily reflect poor emotion understanding, or are these poor results, to a significant extent, a consequence of the single–ground-truth paradigm used in existing benchmarks, one that does not account for the fact that, in many cases, multiple emotions are highly plausible due to ambiguity at different levels?

To address this question, we first conduct a systematic zero-shot evaluation of LLMs on multiple ERC benchmarks using their original annotations. We then re-annotate representative subsets of these datasets with additional human annotators and analyze inter-annotator agreement and disagreement, treating the original dataset label as an additional annotation. This helps us see how LLMs perform when humans agree versus when they do not. It allows us to distinguish clear-consensus cases from those that admit multiple plausible interpretations.

In many existing ERC datasets, our findings suggest that ambiguity allows the same instance to reasonably support different emotion labels. To enable fair evaluation and comparison at scale, this can be addressed either by augmenting datasets to capture such uncertainty or by using an evaluator that tests whether a candidate label is valid given the utterance and its context. We pursue the latter via an LLM-as-Judge that, given a label, an utterance, and dialogue context, predicts label validity; comparison with human annotations yields encouraging results, indicating promise for scaling more reliable evaluation under ambiguity.

Our main contributions are summarized as follows:

- We present a systematic zero-shot evaluation of multiple LLMs across MELD, EmoryNLP, and DailyDialog using full conversational context.
- We re-annotate representative subsets of these datasets and analyse inter-annotater agreement, showing that strong consensus is limited across a substantial portion of ERC instances.
- We evaluate how well LLMs predict emotions when human annotations are treated as the reference labels, and analyze how performance varies across different levels of inter-annotator agreement.
- We investigate an LLM-as-Judge framework as a complementary evaluation tool for assessing emotion plausibility beyond single-label annotations.

2 Related Work

Emotion Recognition in Conversations (ERC) has evolved from early feature-based and neural approaches to more recent large language model (LLM)-based methods. Initial work focused on modeling temporal dynamics and inter-speaker interactions using recurrent and graph-based architectures (Poria et al., 2017; Hazarika et al., 2018), while later studies incorporated external knowledge sources, such as commonsense reasoning and psychological features, to better capture contextual and affective dependencies in dialogue (Zhong et al., 2019; Li et al., 2021).

Despite these methodological advances, ERC benchmarks remain challenging due to intrinsic dataset properties. Most datasets are highly imbalanced, with neutral emotions dominating and minority emotions sparsely represented (Poria et al., 2019; Yang et al., 2022). Moreover, emotion recognition is typically framed as a single-label classification task, implicitly assuming a unique and unambiguous gold label for each utterance.

Recent LLM-based approaches have demonstrated promising performance on ERC tasks. Methods such as CoE (Shen et al., 2025) leverage auxiliary information, including scene descriptions and persona cues, to improve accuracy. However, these gains often rely on dataset-specific contextual signals and do not generalize well across benchmarks (Fu et al., 2024). In parallel, the emergence of large-scale LLMs such as LLaMA (Touvron et al., 2023), Mistral (Jiang et al., 2023), Qwen (Bai et al., 2023), Gemma (Mesnard et al., 2024), and GPT-OSS (Patil et al., 2025) has reshaped ERC research, with many studies relying on fine-tuning

or few-shot prompting tailored to specific benchmarks (Zhang et al., 2023).

However, comparatively little attention has been paid to zero-shot ERC evaluation under full conversational context, where models are not adapted to dataset-specific annotations (Winata et al., 2021). Moreover, prior evaluations largely overlook the role of inter-annotator agreement, making it difficult to disentangle genuine model limitations from ambiguity inherent in emotion annotation. More importantly, existing ERC benchmarks and evaluation protocols rarely consider the possibility that an utterance may admit multiple plausible emotional interpretations, instead enforcing a single-label assumption that overlooks the subjective and context-dependent nature of emotion perception.

3 Experimental Setup

3.1 Datasets

We evaluate four widely used ERC benchmarks that differ significantly in genre and class distribution: MELD (Poria et al., 2019), EmoryNLP (Zahiri and Choi, 2018), DailyDialog (Li et al., 2017), and IEMOCAP (Busso et al., 2008). MELD and EmoryNLP are both derived from the *Friends* TV series; however, MELD uses seven standard emotions and is ∼48% neutral, while EmoryNLP employs a finer-grained, context-dependent inventory. DailyDialog consists of everyday two-speaker exchanges and exhibits a severe class imbalance, with over 80% of utterances labeled as neutral. Finally, IEMOCAP contains scripted and improvised dyadic dialogues, from which they utilize six categorical labels to capture both acted and spontaneous emotional expressions. A comprehensive overview of dataset statistics is provided in Appendix A.1.

3.2 Prompting Protocol

We evaluate each target utterance using a contextual prompting setup, where all preceding dialogue turns with speaker attribution are included to preserve conversational history. This follows prior ERC work showing that dialogue context and speaker states improve emotion recognition (Majumder et al., 2019; Ghosal et al., 2019) and recent studies indicating similar gains for LLMs in zero-shot settings (Qin et al., 2023; Liang et al., 2023). We therefore adopt the full-context setup to assess each model's intrinsic contextual reasoning ability. Decoding is deterministic (temperature = 0.3) to ensure consistency and comparability across models. Further details appear in Appendix A.3.

3.3 Evaluation Metrics

We report three complementary metrics. Accuracy provides an overall measure of correctness but can be dominated by frequent classes in highly imbalanced datasets. Weighted F1 (W-F1) reflects overall performance while accounting for class imbalance, but may still obscure errors on minority emotions. Macro F1 (M-F1) assigns equal weight to all classes, making it essential in ERC where rare emotions often carry crucial affective signals.

4 Limitations in Current Benchmarking

ERC evaluation faces several structural and methodological limitations. Most used datasets are domain-specific or scripted (e.g., MELD and EmoryNLP derived from *Friends*, and DailyDialog collected via crowd-sourced text), which limits ecological validity and encourages models to exploit stylistic or dataset-specific cues rather than develop generalizable affective reasoning.

Benchmarking practices further rely heavily on aggregate metrics such as Accuracy or Weighted-F1. Many recent works, including CFN-ESA (Li et al., 2023a), PFA-ERC (Khule et al., 2024), and InstructERC (Lei et al., 2023), report only overall scores without detailed analysis of minority-class performance. As a result, systematic errors on underrepresented emotions such as *fear*, *disgust*, or *surprise* are often obscured.

In addition, standard ERC evaluation protocols assume a single ground-truth label per utterance. By relying on this setup, evaluation does not distinguish between genuine model errors and cases where multiple emotion interpretations are plausible.

Finally, most prior studies evaluate a single model on a single dataset, limiting insights into cross-model robustness and dataset transferability.

Motivated by these limitations, we now examine zero-shot ERC performance of LLMs under original single-label evaluation.

5 Zero-Shot ERC Performance on Original Annotations

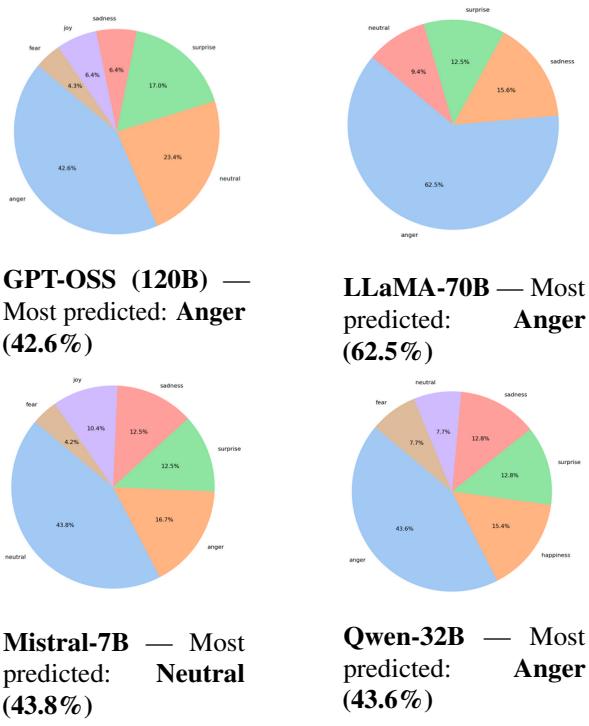
In this section, we present zero-shot results of four LLMs (LLaMA-70B, Qwen-32B, GPT-3.5, Mistral-7B) on three ERC benchmarks (MELD,

Emotion / Metric	LLaMA-70B	Qwen-32B	GPT-OSS (120B)	Mistral-7B
MELD				
Fear	0.333	0.268	0.303	0.250
Disgust	0.398	0.389	0.368	0.244
Macro-F1	0.520	0.502	0.434	0.421
Weighted-F1	0.628	0.616	0.606	0.564
EmoryNLP				
Sad	0.362	0.392	0.335	0.332
Powerful	0.110	0.114	0.097	0.048
Macro-F1	0.274	0.353	0.228	0.239
Weighted-F1	0.355	0.389	0.372	0.315
DailyDialog				
Fear	0.158	0.129	0.121	0.117
Disgust	0.226	0.261	0.284	0.201
Macro-F1	0.396	0.410	0.455	0.323
Weighted-F1	0.746	0.769	0.724	0.633

Table 1: Per-class F1 for minority emotions across datasets and models. Weighted-F1 values are driven by frequent classes such as *Neutral* and *Anger*.

EmoryNLP, DailyDialog), analyzing both metrics and failure patterns to reveal hidden limitations.

5.1 Minority Emotions Analysis



GPT-OSS (120B) — Most predicted: **Anger (42.6%)**

LLaMA-70B — Most predicted: **Anger (62.5%)**

Mistral-7B — Most predicted: **Neutral (43.8%)**

Qwen-32B — Most predicted: **Anger (43.6%)**

Figure 1: Predicted emotion distributions for **Disgust** utterances in MELD across all models.

To better understand model behaviour, we focus on minority emotions because models achieve good performance on majority emotion categories and aagregate metrics tend to mask performance differences on minority emotions. Table 1 reports per-class F1 for minority emotions, while Figure 1 visualizes predicted emotion distributions for *Disgust* in MELD. Across all models, *Disgust* is frequently misclassified as **Anger** or **Neutral**, showing a systematic bias toward high-frequency or semantically related classes(Figure 8). Weighted-F1 is higher in MELD and DailyDialog (0.60–0.75) but drops for EmoryNLP (0.31–0.39), while Macro-F1 remains low, indicating that rare emotions such as *Fear* and *Disgust* are not reliably recognized in zero-shot settings.

Dataset-level interpretation suggests that MELD and DailyDialog contain clearer and more repetitive emotional categories, making them easier for models to classify. In contrast, EmoryNLP includes subtle, context-dependent emotions like *Powerful* and *Peaceful*, which are inherently more difficult to detect (Table 1, Table 14). Feature-based analysis in Appendix A.7 shows that misclassifications are more frequent in utterances containing surface-level cues, such as exclamations, negations, or short replies. These patterns are consistent across datasets and models, indicating that the errors are systematic and largely influenced by dataset characteristics rather than specific model architectures.

5.2 Consistent Failure Modes in Utterances

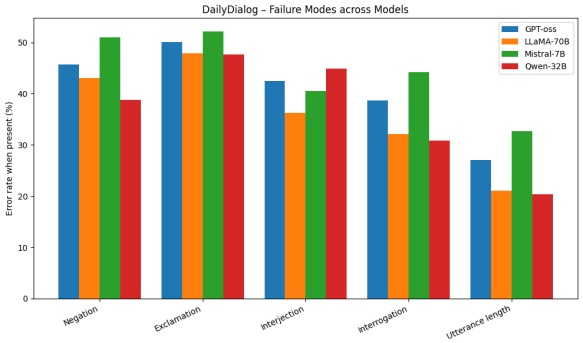


Figure 2: Comparison of failure-mode error rates in DailyDialog across four LLMs (GPT-OSS-120B, LLaMA-70B, Mistral-7B, Qwen-32B). Negations and exclamations show the highest error rates, followed by interjections and interrogations; fewer errors for short replies.

To identify recurring linguistic weaknesses, we first manually inspected 300 misclassified utterances (100 per dataset). This initial review revealed four dominant sources of confusion: negations, exclamations, interjections, and short or minimal replies, which together explained most errors. We then extended this analysis automatically across the entire test sets using a simple rule-based tagging system. For each utterance, we automatically detected whether it contained a negation word (e.g., "not", "don't"), an exclamation mark, a common interjection (e.g., "ugh", "oh", "hmm"), or a question mark, and whether it was shorter than five words. By cross-referencing these tags with model predic-

tions, we computed how often each linguistic pattern coincided with a misclassification. The code, along with all the provided materials, is submitted.

Figure 2 summarizes the results for DailyDialog. Across all models, negation and exclamation consistently produce the highest error rates (around 45–50%), followed by interjections and interrogations, while short utterances cause fewer errors. Under the assumption that each utterance is associated with a single emotion label , the consistency of these results across all four LLMs indicates a shared bias: models rely on surface cues rather than deeper contextual understanding. This accounts for the frequent polarity reversals (e.g., interpreting "I'm not mad" as *Anger*) and the tendency to misclassify neutral or subtle emotions as highly expressive or intense.

However, manual inspection of a subset of these cases indicates that multiple emotional interpretations may be plausible given the context, the intent, the exisiting lexical cues and punctuations. This observation motivates an analysis of human annotation agrrement in ERC benchmarks.

6 Annotation Agreement and Dataset Ambiguity

To better understand whether the misclassification errors observed in the previous section stem from genuine model limitations or from ambiguity in utterances and their context We run a small-scale annotation task of emotions on the different datasets and we analyze human agreement. Specifically, we investigate how often annotators agree versus diverge in emotion labeling and assess the extent to which ERC benchmarks exhibit intrinsic ambiguity, potentially accounting for these differences.

6.1 Re-Annotation Setup and Agreement Measurement

Since ERC benchmarks rely on a single emotion label per utterance, we conducted a controlled re-annotation study on representative subsets of each dataset. For MELD, EmoryNLP, DailyDialog, and IEMOCAP, we randomly sampled dialogue turns by selecting the first four turns from each conversation until 100 turns were obtained per dataset. The sampling process was designed to ensure that all emotion categories present in each dataset were included, allowing annotation agreement to be examined across the full label space. Each selected turn was annotated by four additional human annotators, who were instructed to assign a *single emotion label* chosen from the original emotion set of the corresponding dataset. The original dataset annotation was treated as an additional annotator, resulting in five annotations per turn. Inter-annotator agreement was assessed using Cohen's κ, computed pairwise and averaged per dataset. The resulting agreement scores are consistently low across benchmarks, with $\kappa = 0.31$ for MELD, 0.29 for EmoryNLP, 0.28 for DailyDialog, and 0.24 for IEMOCAP. These values are comparable to, and in some cases lower than, those reported in the original dataset papers (Poria et al., 2019; Zahiri and Choi, 2018; Li et al., 2017; Busso et al., 2008), confirming that annotation ambiguity is an inherent characteristic of ERC data rather than an artifact of our annotation protocol. Beyond that, we analyze agreement at the instance level by grouping turns according to the number of annotators assigning the same emotion. We define five agreement regimes: *5/5* (full agreement), *4/5* (strong agreement), *3/5* (moderate agreement), *2/5* (low agreement), and *1/5* (complete disagreement). Figure 3 presents agreement histograms for MELD and EmoryNLP, with bars further decomposed by emotion category.

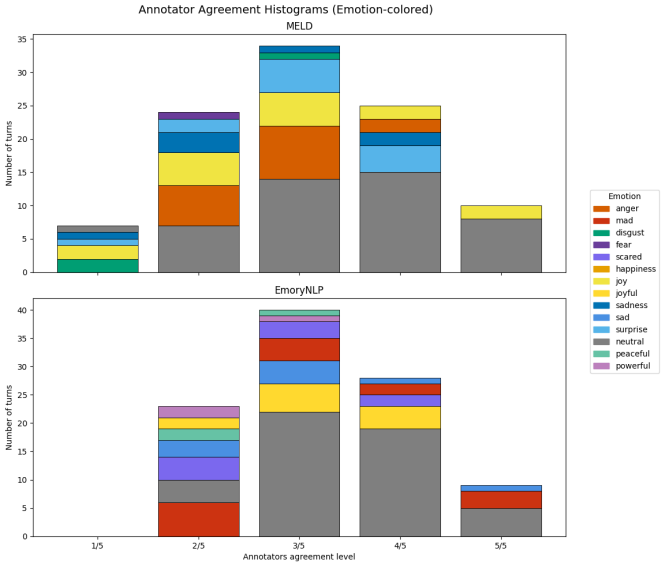


Figure 3: Annotator agreement distributions for MELD (top) and EmoryNLP (down), stratified by emotion category. High-consensus cases (4/5 and 5/5) are dominated by the neutral class, while minority emotions concentrate in intermediate agreement regimes (2/5 and 3/5).

The histograms reveal that high-consensus cases (4/5 and 5/5) account for only a minority of the annotated turns, and are largely dominated by the *neutral* emotion. In contrast, minority emotions such

as *fear*, *disgust*, *sadness*, *peaceful*, and *powerful* appear predominantly in the 1/5, 2/5 and 3/5 agreement regimes. This pattern indicates that these emotions are not only underrepresented in the datasets, but also intrinsically more ambiguous for annotators. Rather than reflecting random disagreement, this indicates that an utterance can support multiple plausible emotional interpretations. These findings highlight a structural limitation of single-label ERC benchmarks: many instances lack a clear emotional consensus even among humans. Agreement and emotion distributions for other datasets follow similar trends and are reported in Appendix A.11.

6.2 Linguistic and Contextual Factors Driving Annotation Disagreement

To analyze the sources of annotation disagreement, we manually label 100 turns per dataset using five qualitative cues: *Intent Clarity*, which distinguishes between clear and ambiguous communicative purposes; *Intent Count*, identifying whether an utterance conveys a single intent or multiple ones (e.g., informing, requesting information, warning, encouraging, complaining, agreeing, mocking, or reproaching); *Context Clarity*, assessing if the surrounding dialogue provides sufficient information to resolve the emotion; *Lexical Cue Presence*, noting the existence of explicit affective terms; and *Punctuation*, which categorizes markers like "!!!" or "???" that amplify intensity. Together, these cues allow us to identify how ambiguity arises from sparse context, overlapping intents, and the reliance on subjective inference in the absence of explicit lexical. To further validate these observations, we train a decision tree classifier using the annotated cues as features and agreement level as target.

Dataset	Acc.	Prec.	Rec.	F1
MELD	0.88	0.81	0.88	0.84
DailyDialog	0.80	0.84	0.80	0.74
EmoryNLP	0.52	0.83	0.52	0.60
IEMOCAP	0.72	0.83	0.72	0.70

Table 2: Decision tree performance for predicting annotation agreement levels using linguistic and contextual cues. Agreement is modeled using three classes: low agreement (1/5), medium agreement (2/5-3-5) and high agreement (4/5-5/5). The decision tree is trained on 75% of the annotated data and evaluated on the remaining 25%.

Table 2 reports the performance of a decision tree classifier trained to predict agreement levels using the five annotated linguistic and contextual cues. Across datasets, the model achieves strong performance on MELD (accuracy 0.88) and DailyDialog (accuracy 0.80), indicating that a small set of interpretable cues is sufficient to explain a large portion of annotation variability. Performance on IEMOCAP remains moderate (accuracy 0.72), while EmoryNLP exhibits substantially lower accuracy (0.52), reflecting the finer-grained and more context-dependent emotion taxonomy of the dataset. Overall, these results confirm that annotation disagreement is not random.

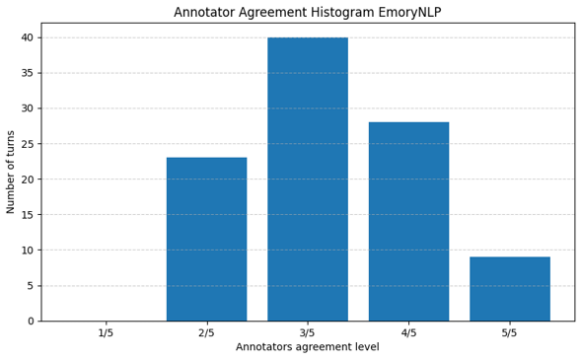


Figure 4: Annotator agreement distribution for EmoryNLP. Most turns fall into intermediate agreement regimes (2/5 and 3/5), indicating substantial annotation ambiguity.

We begin by examining the overall distribution of annotator agreement levels. Figure 4 shows the agreement histogram for EmoryNLP, where the majority of turns fall into intermediate agreement regimes. In particular, cases with 2/5 and 3/5 agreement dominate the distribution, while full consensus (5/5) remains relatively rare. To better understand the linguistic and contextual properties underlying these ambiguous cases, we focus specifically on turns in 2/5 and 3/5 agreement. For these turns, we analyze the distribution of the manually annotated cues capturing the characteristics of the utterances. Figure 5 presents the cue distributions for ambiguous instances.

Several consistent patterns across datasets emerge from this analysis. First, ambiguous turns are strongly associated with *unclear intent*, suggesting that annotators struggle to infer a single dominant communicative goal. Conversely, utterances with high agreement (4/5 and 5/5) are predominantly associated with **clear intent**, indicating that annotaters consistently identify a single dominant emotion (see Appendix B Second, utterances expressing multiple competing intents (2+) occur more frequently than single-intent utterances, indicating that cause is a key driver of disagree-

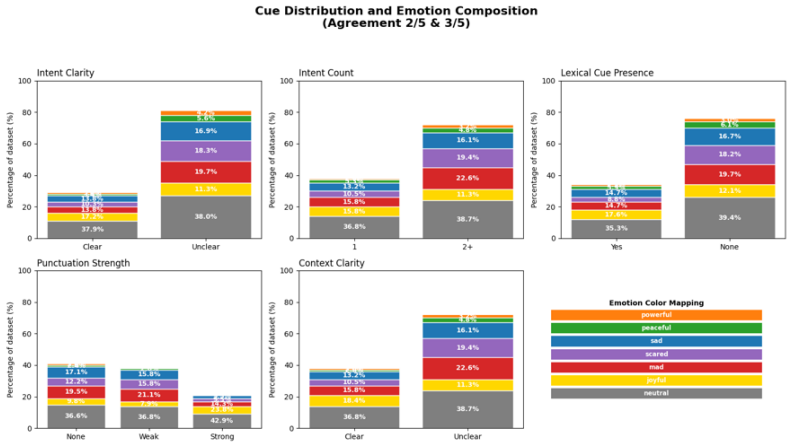


Figure 5: Distribution of linguistic and contextual cues for turns with 2/5 and 3/5 agreement in EmoryNLP. Ambiguous instances are predominantly associated with unclear intent, multiple competing intents, absence of lexical emotion cues, unclear context, and weak or strong punctuation.

ment. In contrast, utterance with high agreement are typically characterized by a single dominant intent. Third, the absence of explicit emotion lexical cues is prevalent among 2/5 and 3/5 cases, forcing annotators to rely on contextual inference rather than affective markers. Context clarity further contributes to ambiguity. Turns annotated as having *unclear context* are overrepresented and are often associated with minority emotion categories across datasets.

Overall, this analysis demonstrates that high ambiguity (multiple intents in an utterance, unclear intent, unclear context, and the absence of lexical affective cues) leads to multiple possible valid interpretations, whereas the opposite would lead towards consensus. These findings motivate considering the evaluation of ERC models on existing benchmarks beyond single-label classification, allowing for different emotion perspectives.

7 LLM Agreement under Multi-Annotated Emotion Scenarios

To better reflect the intrinsic ambiguity of conversational emotion recognition, we evaluate LLM predictions under a multi-annotation–aware setting, where a prediction is considered correct if it matches at least one emotion assigned by human annotators. We evaluate eight LLMs of varying sizes and architectural properties in a zero-shot setting with full conversational context across ERC benchmarks. Table 3 summarizes the resulting accuracies. Across datasets, we observe substantial convergence in model predictions, with inter-model agreement reaching 72% on MELD, 62% on DailyDialog, 56% on IEMOCAP, and 68% on EmoryNLP. This convergence indicates that, despite architectural diversity, LLMs tend to favor similar emotional interpretations when multiple plausible labels are allowed. However, agreement is strongly emotion-dependent. Across all datasets, more than 80% of agreed predictions correspond to *neutral* or positive emotions such as *joy* whereas minority emotions including *disgust*, *fear*, *powerful*, and *peaceful* remain consistently under-predicted. This pattern mirrors the human annotation distributions observed in Section 5.2. Model scale provides partial robustness but does not fundamentally resolve these biases. Larger models (e.g., LLaMA-70B, GPT-OSS-120B, DeepSeek-R1-70B) exhibit more stable performance across datasets. Interestingly, models explicitly optimized for reasoning do not consistently outperform non-reasoning LLMs in ERC. Overall, these findings motivate evaluation beyond strict single-label emotion.

8 Evaluation with LLM-as-Judge

To further investigate the plausibility of emotions in ERC benchmarks, we introduce an *LLM-as-Judge* framework that evaluates the semantic compatibility of human emotion annotations. The LLM is not used to assess ERC model predictions; instead, it acts as a semantic judge that determines whether a human-assigned emotion label is plausible given the conversational context. For each utterance, we use **Gemma-3-27B** as the judge. The model is provided with the full dialogue context and the dataset emotion labels, and selects those it considers plausible. A human annotation is deemed *compatible*

Dataset	LLaMA 70B	GPT-OSS	Qwen 32B	Mistral 7B	Gemma 27B	DeepSeek-R1	LLaMA 8B	GPT-OSS Safe	*Inter-model agreement on predicted emotion (%)*	*Agreement on neutral and positive emotions (%)*
MELD	0.81	0.87	0.89	0.91	0.94	0.88	0.82	0.93	72	85
DailyDialog	0.93	0.94	0.98	0.91	0.86	0.94	0.84	0.91	62	88
EmoryNLP	0.83	0.82	0.93	0.88	0.91	0.83	0.76	0.90	68	82
IEMOCAP	0.84	0.79	0.80	0.73	0.86	0.83	0.72	0.83	56	79

Table 3: Multi-annotation–aware accuracy of eight LLMs across ERC benchmarks. A prediction is considered correct if it matches at least one human-annotated emotion. Models exhibit strong inter-model agreement, largely driven by neutral and positive emotions, while minority emotions remain challenging across datasets.

if its assigned emotion is among the selected labels. This formulation defines a binary agreement task over (utterance, emotion) pairs, enabling evaluation beyond exact label matching. We select Gemma-3-27B due to its stable zero-shot behavior across ERC datasets.

Dataset	Acc.	Rec.	F1	FPR	N
DailyDialog	0.81	0.87	0.73	0.22	700
MELD	0.73	0.84	0.70	0.33	700
IEMOCAP	0.70	0.85	0.70	0.41	600
EmoryNLP	0.72	0.76	0.65	0.30	700

Table 4: Agreement between LLM-as-Judge and human annotations. Metrics quantify semantic compatibility; N denotes the number of evaluated (utterance, emotion) pairs.

Table 4 shows that the LLM-as-Judge achieves consistently good performances across datasets, with accuracy ranging from 0.70 to 0.81 and F1 scores between 0.65 and 0.73. High recall values (0.76–0.87) indicate that the judge reliably recognizes emotions considered plausible by human annotators, even in ambiguous conversational contexts.

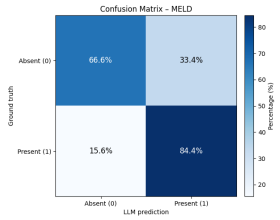


Figure 6: Confusion matrix for LLM-as-Judge agreement on MELD.

Agreement Level	Count	Percentage (%)
1/5 (No agreement)	115	78.77
2/5	24	16.44
3/5	4	2.74
4/5	1	0.68
5/5	2	1.37

Table 5: Distribution of LLM-as-Judge false positives by human agreement level in MELD.

Figure 6 shows the confusion matrices for the LLM-as-Judge on MELD. False positives occur in utterances with low inter-annotator agreement, as shown in Table 5: over 95% arise from 1/5 or 2/5 agreement cases and are rare when human consensus is high. This suggests that many incompatibilities identified by the LLM-as-Judge arise in cases where emotional interpretation is less clear.

9 Discussion

Our study shows that ERC presents substantial challenges for single-label evaluation. Re-annotation indicates that high consensus (4/5 agreement) is relatively infrequent, especially for minority emotions such as fear and disgust. Disagreement follows systematic patterns associated with linguistic factors, including unclear intent, competing intentions, and limited contextual grounding. Consistent with this observation, the LLM-as-Judge analysis shows that discrepancies primarily arise in low-consensus cases, while high-agreement instances are rarely disputed. Together, these findings highlight the limitations of single-label, metric-driven ERC evaluation and motivate agreement-aware approaches that better capture the variability of emotional interpretation in dialogue.

10 Conclusion

In this work, we presented a comprehensive analysis of Emotion Recognition in Conversations under a zero-shot setting, highlighting limitations of ERC benchmarks and evaluation practices. By combining large-scale LLM evaluation, re-annotation and agreement analysis, cue-based investigation, and an LLM-as-Judge framework, we show that a substantial portion of ERC data is not well captured by strict single-label annotations. Our results show that many apparent model errors reflect genuine uncertainty. These findings suggest that future ERC datasets should move from single to multiple plausible emotions per utterance through multi-label annotations, or agreement-aware evaluation protocols.

11 Limitations

This study has several limitations. First, annotators are still asked to select a single dominant emotion from predefined label sets, which may underestimate the full range of plausible emotional interpretations. Second, the LLM-as-Judge framework relies on a specific prompting strategy and a fixed set of large language models; alternative formulations or models may lead to different compatibility judgments. Finally, our analysis focuses on text-only ERC benchmarks. Future work should investigate how multimodal information, including audio, visual cues, and prosodic features, interacts with annotation agreement and ambiguity in emotion recognition across diverse conversational settings.

References

Zhicong Bai, Hongyi Bai, Hao Chen, and 1 others. 2023. Qwen: Open foundation models by alibaba cloud. *arXiv preprint arXiv:2309.16609*.

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. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*.

Jing Chen and Ling Xiao. 2024. Sentiment reasoning of llms in dialogue emotion tasks. *Transactions on Affective Computing*.

Nicholas Cummins, Stefan Scherer, Jarek Krajewski, Stefan Schnieder, Julien Epps, and Thomas F Quatieri. 2019. A review of depression and suicide risk assessment using speech analysis. *Speech Communication*, 71:10–49.

Li Fu, Zhenyu Zhang, and Haonan Chen. 2024. Contextual emotion understanding with large language models. *ArXiv preprint arXiv:2402.04122*.

Deepanway Ghosal, Navonil Majumder, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. Cosmic: Commonsense knowledge for emotion identification in conversations. In *Proceedings of ACL*.

Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. 2019. Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation. In *EMNLP*.

Devamanyu Hazarika, Soujanya Poria, Amir Zadeh, Erik Cambria, Louis-Philippe Morency, and Rada Mihalcea. 2018. Conversational memory network for emotion recognition in dyadic dialogue. In *Proceedings of NAACL*.

Albert Q Jiang and 1 others. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Tanmay Khule, Rishabh Agrawal, and Apurva Narayan. 2024. Pfa-erc: Pseudo-future augmented dynamic emotion recognition in conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16196–16207.

Ming Lei, Hao Wang, Peixiang Zhong, and Rui Chen. 2023. Instructerc: Instruction-based emotion recognition in conversation. In *Findings of ACL*.

Jiang Li, Xiaoping Wang, Yingjian Liu, and Zhigang Zeng. 2023a. Cfn-esa: A cross-modal fusion network with emotion-shift awareness for dialogue emotion recognition. *arXiv preprint*.

Jie Li, Jian Zhang, and Dongmei Xu. 2021. Incorporating psychological knowledge for emotion recognition in conversations. In *Proceedings of EMNLP*.

Xiaolong Li, Weiqi Wang, and et al. 2023b. Multimodal depression estimation in the wild: A challenge dataset and baseline results. In *Proceedings of ICMI*.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Jie Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of IJCNLP*.

Wei Liang, Yi Chen, and Min Huang. 2023. Zero-shot emotion recognition in conversations via prompt-based llms. In *Proceedings of EMNLP*.

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. 2019. Dialoguernn: An attentive rnn for emotion detection in conversations. In *AAAI*.

Thomas Mesnard and 1 others. 2024. Gemma: Open models based on gemini research and technology. Technical report, Google DeepMind.

Rakesh Patil, Zhiyu Shen, Yanghui Rao, and DeepSeek-AI Team. 2025. Gpt-oss: Open-source large reasoning models approaching gpt-4 performance. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Prateek Vij, Navonil Majumder, and Alexander Gelbukh. 2017. Context-dependent sentiment analysis in user-generated videos. In *ACL*.

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 ACL*.

Xueying Qin, Qian Wang, and Jianshu Li. 2023. Llm-based emotion recognition in conversations: Contextual prompts and zero-shot evaluation. In *Findings of ACL*.

Hannah Rashkin, Eric Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of ACL*.

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

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Akshita Bhargava, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Genta Indra Winata, Andrea Madotto, Zhaojiang Lin, and Pascale Fung. 2021. Multilingual zero-shot emotion recognition with pre-trained language models. In *Proceedings of ACL*.

Xu Yang, Min Zhang, and Pengjie Li. 2022. Contrastive representation learning for emotion recognition in conversation. In *COLING*.

Sayyed M Zahiri and Jinho D Choi. 2018. Emotion detection on tv show transcripts with sequence-based convolutional neural networks. In *AAAI Workshop on Affective Content Analysis*.

Wei Zhang, Yu Liu, and Xin Tang. 2023. Emotion-llm: Fine-tuning large language models for emotion understanding in conversations. In *Proceedings of ACL*.

Peixiang Zhong, Di Zhang, and Hao Wang. 2019. Knowledge-enriched transformer for emotion detection in conversations. In *EMNLP*.

Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2018. Emotional chatting machine: Emotional conversation generation with internal and external memory. In *AAAI*.

A Appendix

A.1 Datasets & Emotions distribution

Dataset	Conversations			Utterances			$\lvert\mathcal{E}\rvert$	Neutral (%)	Imbalance Ratio
	Train	Val	Test	Train	Val	Test			
EmoryNLP	713	99	85	9934	1344	1328	7	29.95	4:1
MELD	1038	114	280	9989	1109	2610	7	48.21	18:1
DailyDialog	11118	1000	1000	87170	8069	7740	7	83.24	1156:1

Table 6: Dataset statistics for ERC benchmarks used in this study. $\mathcal{E}$ denotes the number of emotion classes. Imbalance ratio and Neutral proportion illustrate dataset skew.

We conducted our experiments on four widely used Emotion Recognition in Conversation (ERC) benchmarks: DailyDialog, IEMOCAP MELD, and EmoryNLP. The figures below present the distribution of emotions within MELD and EmoryNLP in the train Set. As can be observed, all three benchmarks exhibit a highly imbalanced class distribution, with certain emotions such as Neutral or Joy dominating the datasets, while minority classes like Fear, Disgust, or Powerful appear only sparsely. This imbalance introduces significant challenges for ERC models, as high overall accuracy or weighted F1 scores may obscure systematic weaknesses in recognizing these underrepresented emotions.

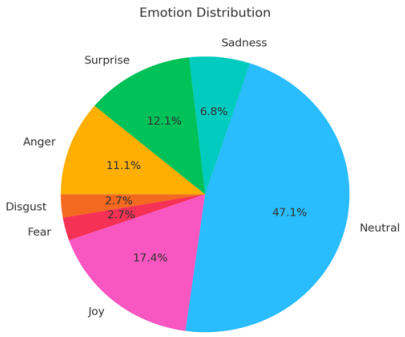


Figure 7: Emotion distributions Train Set of MELD

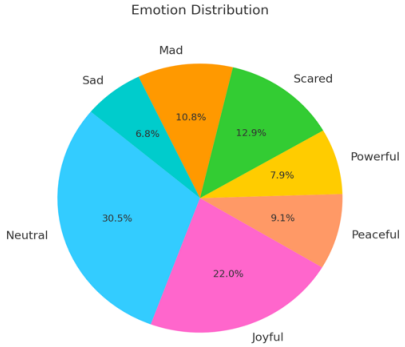


Figure 8: Emotion distributions Train Set of EmoryNLP.

A.2 Implementation Details

We ran **Mistral-7B** in FP16 precision, while **LLaMA-70B**, **Qwen-32B**, and **GPT-OSS-120B** were evaluated in quantized settings 4-bit. All decoding was performed with **temperature = 0.2** to ensure deterministic outputs across runs. This configuration avoids randomness in generations and guarantees reproducibility of the reported results.

A.3 Prompt Structure

We experimented with multiple prompting formulations before converging on the final version shown in Table 7. Early prompt variants used generic or few-shot instructions (e.g., "Classify the emotion of this sentence") but produced inconsistent or verbose outputs. Models frequently inferred emotions from the interlocutor's perspective, overused the *Neutral* label, or generated multiple labels instead of a single discrete class. After iterative testing across MELD, EmoryNLP, and DailyDialog, we observed that explicitly constraining the task to the *speaker's own emotional state*, structuring the input into **context** and **utterance** segments, and enforcing a strict **JSON output format** yielded the most reliable and reproducible results. This final design reduces ambiguity, ensures consistent syntactic output for automatic parsing, and aligns with prior ERC prompting frameworks emphasizing explicit contextual grounding and role delimitation (Lei et al., 2023; Shen et al., 2025; Fu et al., 2024).

Table 7: Final prompt template for Conversational Emotion Recognition (ERC) after iterative design.

Prompt for Conversational Emotion Recognition (ERC)	
[System]	You are performing **Conversational Emotion Recognition (ERC)**. Use ONLY the utterance and its dialogue context (previous turns). Focus strictly on the **speaker's own emotional state** expressed in the utterance. Do NOT infer emotions based on how others might receive or interpret it.
[Dialogue Context]	All previous turns from the same dialogue are provided below. If there is no prior context, write *(no prior context)*. `(Example)` Monica: "So ah, Phoebe, how was your date?" Phoebe: "Oh well y'know." Monica: "Yeah, I do know." Chandler: "Don't worry."
[Utterance to Classify]	`Phoebe: "God, I hope they kick his ass!"`
[Task]	Classify the utterance into exactly **one emotion**. It can be ONLY one of these emotions: `[joy, sadness, anger, fear, disgust, surprise, neutral]` (or your own emotion set). Do NOT add any other emotions. Provide a short grounded reasoning (cite specific words/phrases from the utterance or dialogue context).
[Output Format]	Return **STRICT JSON ONLY** in the following format: `{ "emotion": "anger", "reason": "Aggressive tone and verb 'kick his ass' express anger." }`

A.4 Analysis Protocol

To systematically investigate model errors beyond aggregate metrics, we designed a two-stage analysis protocol combining *visual inspection* and *quantitative validation*. The process was implemented in Python through three dedicated scripts.

Visual inspection. First, we enriched the prediction outputs by color-coding true and predicted emotions using a consistent palette (one color per class). A script automatically compared the gold and predicted labels and highlighted mismatches in Excel files, also alternating background shades across dialogues for readability. This produced a visually structured dataset that facilitated rapid spotting of misclassifications and utterance-level anomalies.

Error distribution. A second script computed classical diagnostic statistics. It generated confusion matrices, per-class counts of true vs. predicted labels, and classification reports including precision, recall, and F1-scores. These outputs provided a quantitative overview of systematic confusions between emotions and highlighted class imbalance effects.

Feature-based error validation. Building on insights from the first two steps, we developed a third script to annotate utterances with binary indicators of error-prone features (e.g., presence of negation, interjections, punctuation, utterance length). The script automatically added six feature columns to the dataset and flagged whether each utterance exhibited the characteristic. Error rates were then recomputed conditionally on these features, both globally and per class, yielding detailed tables of error ratios. Pie charts summarizing error distributions were also generated and compiled into an HTML dashboard.

Outcome. This protocol enabled both qualitative exploration and quantitative validation. The visual stage facilitated hypothesis generation about recurring weaknesses, while the feature-based quantification confirmed their prevalence and provided evidence for systematic error categories. The resulting annotated datasets, confusion matrices, and HTML reports served as the foundation for the error taxonomy presented in Section 5.2(Consistent Failure Modes in Utterances).

A.5 Re-annotation Protocol.

To analyze annotation ambiguity in ERC benchmarks, we conducted a targeted re-annotation of a limited, representative subset of the data. For each dataset (MELD, EmoryNLP, DailyDialog, and IEMOCAP), we selected 100 dialogue turns covering all available emotion categories. Each turn was independently annotated by four additional annotators, who were instructed to assign a single emotion label chosen exclusively from the original dataset's emotion inventory, based on the full conversational context. The original dataset annotation was treated as an additional reference label, yielding five annotations per turn. No personal or sensitive information was collected, and the task involved only the interpretation of pre-existing, publicly available text. This controlled re-annotation setup was designed solely to quantify agreement patterns and assess annotation variability, without altering or extending the original datasets.

A.6 Additional Results: Confusion Matrices

In this section, we provide the complete confusion matrices for all models for **MELD**. These figures complement the representative examples shown in the main paper (Section 5). They confirm the systematic blind spots we described, especially the dominance of *Neutral* predictions and the collapse of minority emotions such as *Fear*, *Disgust*, or *Powerful*. Across all models, minority categories are systematically absorbed into high-frequency or semantically adjacent classes (e.g., *Fear* → *Neutral*, *Disgust* → *Anger*), revealing that aggregate scores are largely driven by majority-class precision.

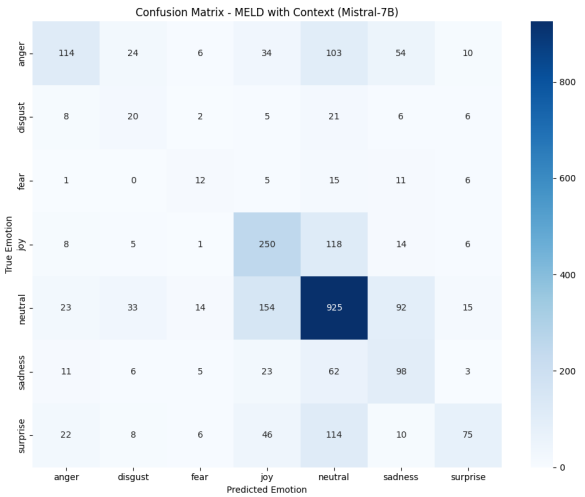


Figure 9: **MELD – GPT-OSS-120B. Fear** and **Disgust** are almost never recognized correctly and are largely predicted as **Neutral** or **Anger**, indicating strong attraction to dominant emotion categories. **Joy** and **Surprise** also show overlap, reflecting surface-form confusion from exclamatory expressions. The strong diagonal for **Neutral** inflates weighted-F1 despite weak coverage of minority classes.

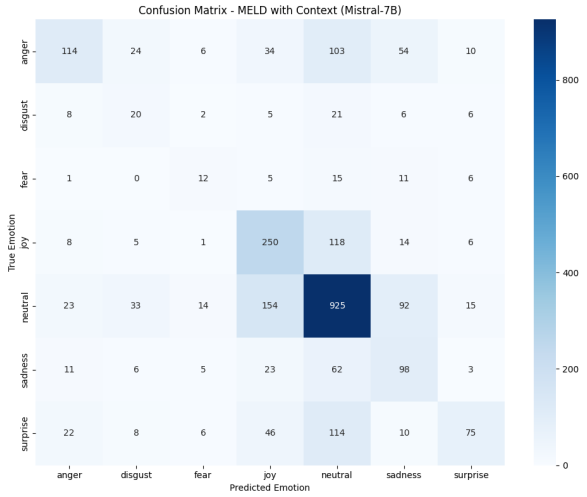


Figure 10: **MELD – Mistral-7B.** Minority emotions such as **Fear** and **Disgust** collapse almost entirely into **Neutral**, while **Sadness** is frequently misread as **Joy**. This reveals a polarity-bias effect: low-valence emotions are absorbed into higher-valence or neutral predictions. Although **Joy** is better recognized, the imbalance in recall persists across emotional extremes.

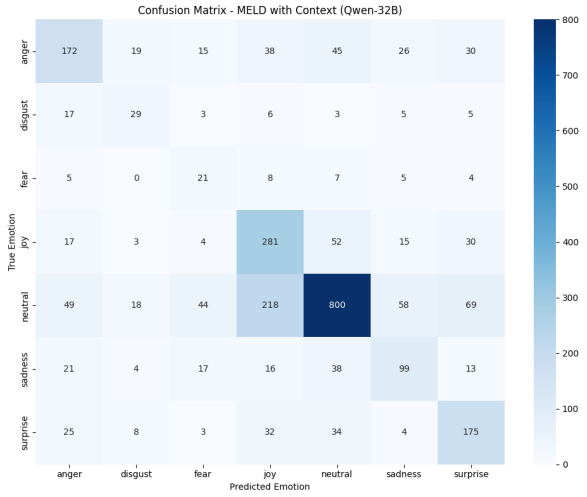


Figure 11: **MELD – Qwen-32B.** While **Joy** and **Neutral** are more distinct, rare classes (**Fear**, **Disgust**) continue to be absorbed by **Anger** or **Neutral**. **Sadness** occasionally shifts toward **Joy**, suggesting confusion in valence polarity rather than arousal intensity. These errors reinforce that higher model scale does not guarantee balanced emotional discrimination.

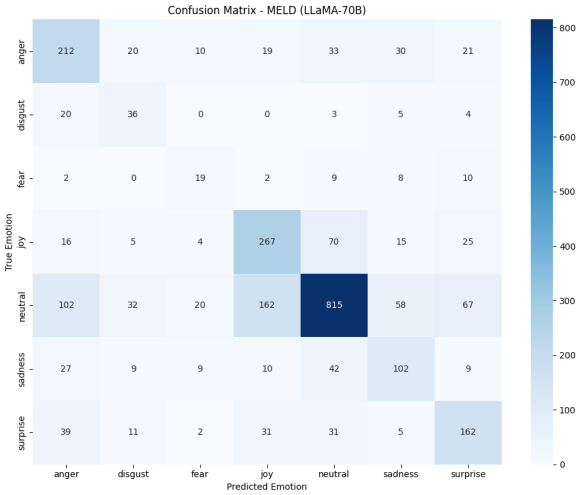


Figure 12: **MELD – LLaMA-70B.** LLaMA exhibits the clearest **Neutral** diagonal and relatively stable **Joy–Surprise** boundary, but still confuses **Disgust** and **Fear** with **Anger**. The persistence of this confusion across all models highlights a structural limitation in emotion separation—particularly for low-frequency or context-dependent states.

A.7 Feature-Based Error Analyses

The following figures show the percentage of presence for various sources of models misclassification across the test sets of DailyDialog, EmoryNLP, and MELDs datasets. These analyses provide insight into the distribution of key features, highlighting their varying frequencies and their potential impact on emotion recognition performance.

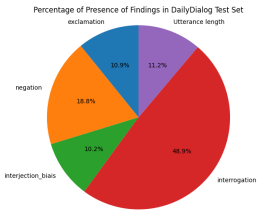


Figure 13: DailyDialog

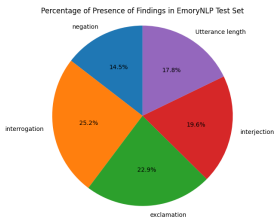


Figure 14: EmoryNLP

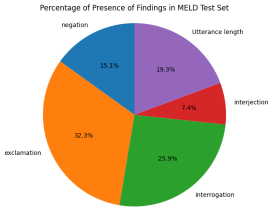


Figure 15: MELD

Figure 16: Presence of linguistic features (e.g., negation, exclamation, interrogation, short replies) across DailyDialog, EmoryNLP, and MELD.

The figures above show the distribution of various type of errors in the DailyDialog, EmoryNLP, and MELD test sets. In the DailyDialog test set, Interrogation is the most prevalent feature, accounting for 48.9% of the dataset. The EmoryNLP dataset shows a more balanced distribution, with Interrogation and Exclamation as the dominant problem at 25.2% and 22.9%, respectively. In the MELD test set, Exclamation is the most frequent finding, representing 32.3% of the dataset, followed by Interrogation at 25.9%.

These distributions suggest that the DailyDialog dataset is heavily skewed towards conversational elements like Interrogation, while EmoryNLP and MELD contain a broader mix of emotional features, with Exclamation and Interrogation being more equally distributed. Understanding the presence of these features is crucial for building models that effectively generalize across different datasets, as the emotional dynamics and conversational context can vary significantly across datasets.

We report failure mode–conditioned error rates across models for EmoryNLP and DailyDialog. As seen in the figures below, models consistently show higher error rates in the presence of *negations*, *exclamations*, and *short utterances*, while interrogations and interjections remain challenging but slightly less severe. These results confirm that blind spots are tied to structural features of dialogue rather than dataset-specific artifacts.

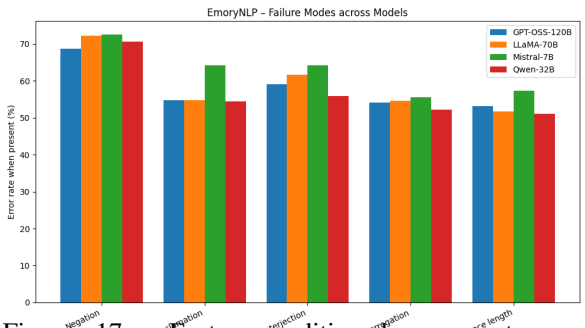


Figure 17: Feature-conditioned error rates on EmoryNLP across GPT-OSS-120B, LLaMA-70B, Mistral-7B, and Qwen-32B.

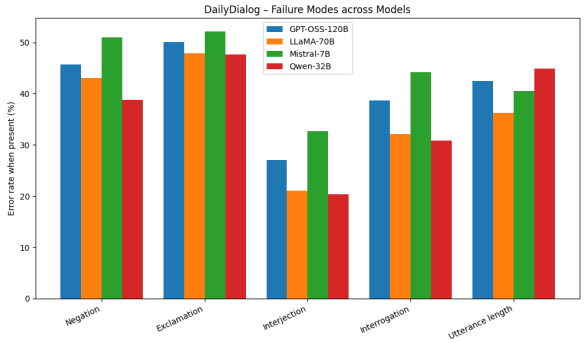


Figure 18: Feature-conditioned error rates on DailyDialog across GPT-OSS-120B, LLaMA-70B, Mistral-7B, and Qwen-32B.

A.8 Illustrative Error Examples

Speaker	Utterance	True Emotion	Predicted Emotion
Rachel Green	So uh, Ryan, were you shipping off to?	**Joyful**	**Neutral**
Ryan	I really can't say.	**Neutral**	**Neutral**
Ross Geller	So do you have like any nuclear weapons on board?	**Joyful**	**Neutral**
Ryan	I can't say.	**Neutral**	**Neutral**

Table 8: Example from **EmoryNLP** using **Mistral-7B**. Type of error: *Question bias* — the interrogative form is misinterpreted as neutral instead of joyful.

Speaker	Utterance	True Emotion	Predicted Emotion
Monica	Oh my God! You got engaged!	**Surprise**	**Joy**
Rachel	Can you believe it?!	**Surprise**	**Joy**
Chandler	Wow! That's... unexpected.	**Surprise**	**Joy**

Table 9: Example from **MELD** using **Qwen-32B**. Type of error: *Exclamation bias* — exclamatory cues ("!", "Oh my God!", "Wow!") trigger **joy** predictions instead of the intended **surprise**.

Speaker	Utterance	True Emotion	Predicted Emotion
Ross	I'm not mad, really.	**Neutral**	**Anger**
Monica	You sound like you are, though.	**Neutral**	**Neutral**
Ross	No, it's fine. I'm just tired.	**Neutral**	**Anger**
Rachel	Okay… I'll stop asking then.	**Sadness**	**Neutral**

Table 10: Example from **DailyDialog** using **LLaMA-70B**. Type of error: *Negation reversal* — the presence of negation ("not mad") triggers polarity inversion, producing **anger** predictions. Despite context indicating calm reassurance, the model relies on surface lexical polarity rather than semantic composition.

Speaker	Utterance	True Emotion	Predicted Emotion
Rachel	Ugh, this day has been awful.	**Disgust**	**Neutral**
Monica	What happened? Did something go wrong?	**Neutral**	**Neutral**
Rachel	Everything. I just want it to end.	**Disgust**	**Sadness**
Chandler	Maybe a coffee will help.	**Neutral**	**Neutral**

Table 11: Example from **MELD** using **Mistral-7B**. Type of error: *Interjection bias* — expressions like "Ugh" signal frustration or disgust, but are flattened to **neutral** or misread as **sadness**. The model fails to interpret non-verbal affective cues typical of spontaneous dialogue.

Speaker	Utterance	True Emotion	Predicted Emotion
Monica	So, did you talk to him?	**Neutral**	**Neutral**
Chandler	Yeah.	**Neutral**	**Joy**
Monica	And? What did he say?	**Neutral**	**Neutral**
Chandler	Nothing special.	**Neutral**	**Sadness**

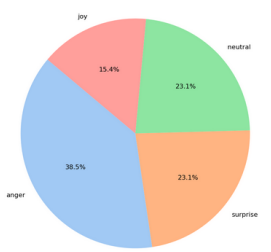
Table 12: Example from **DailyDialog** using **GPT-OSS-120B**. Type of error: *Short reply ambiguity* — minimal responses ("Yeah", "Nothing special") are interpreted inconsistently, oscillating between **joy** and **sadness**. This inconsistency indicates weak sensitivity to discourse pragmatics.

Speaker	Utterance	True Emotion	Predicted Emotion
Phoebe	I had this nightmare last night.	**Fear**	**Neutral**
Rachel	Oh no! What was it about?	**Surprise**	**Surprise**
Phoebe	Everything was falling apart... I couldn't move.	**Fear**	**Sadness**
Monica	That sounds horrible!	**Sadness**	**Neutral**

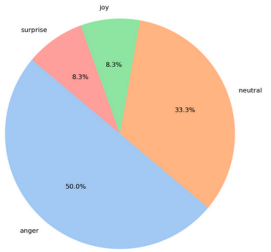
Table 13: Example from **MELD** using **Qwen-32B**. Type of error: *Low-frequency emotion collapse* — the emotion **fear** is consistently misclassified as **sadness** or **neutral**, even with explicit contextual cues ("nightmare", "couldn't move"). This reflects a bias toward majority affective categories.

A.9 Emotion–Feature Interaction Analysis

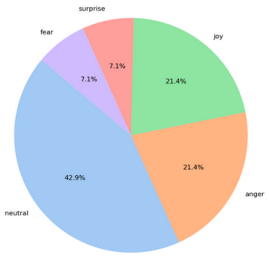
To better understand how surface features affect predictions, we analyzed the interaction between specific emotions and linguistic features (e.g., negation, exclamations, interjections). For each combination, we computed the distribution of predicted labels when the feature was present.



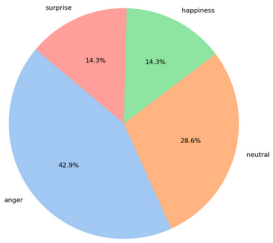
GPT-OSS (120B) — Most predicted: **Anger (38.5%)**



LLaMA-70B — Most predicted: **Anger (50.0%)**

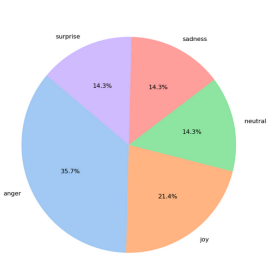


Mistral-7B — Most predicted: **Neutral (42.9%)**

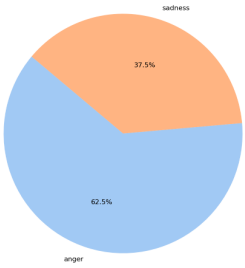


Qwen-32B — Most predicted: **Anger (42.9%)**

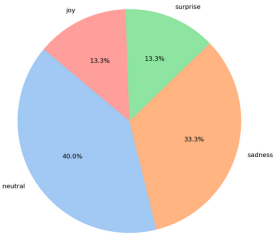
Figure 19: Distribution of predictions for **Sadness** utterances containing exclamations across all four models. While the ground truth emotion is **Sadness**, all models show strong confusion toward high-arousal categories such as **Anger** and **Neutral**. This suggests that the presence of exclamations biases models toward interpreting the emotional tone as more intense or externally directed, leading to systematic misclassification of low-arousal emotions like sadness.



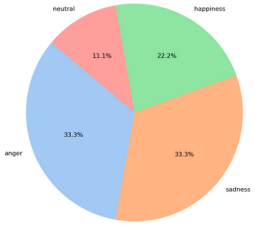
GPT-OSS (120B) — Most predicted: **Anger (35.7%)**



LLaMA-70B — Most predicted: **Anger (62.5%)**



Mistral-7B — Most predicted: **Neutral (40%)**



Qwen-32B — Most predicted: **Anger (33.3%)**

Figure 20: Distribution of predicted emotions for **Disgust** utterances containing interrogations across all four models. Dominant predictions reveal confusion toward frequent classes such as **Anger** and **Neutral**, confirming a consistent feature-level bias.

A.10 Dataset- and Model-Specific Results

We summarize the macro-F1 and weighted-F1 scores per dataset and model, highlighting that blind spots (minority emotions, interjections, negations, short utterances) recur independently of the dataset or model architecture.

Emotion / Metric	LLaMA-70B	Qwen-32B	GPT-OSS (120B)	Mistral-7B
		MELD		
Anger	0.556	0.528	0.542	0.429
Disgust	0.398	0.389	0.368	0.244
Fear	0.333	0.268	0.303	0.250
Joy	0.598	0.000	0.572	0.544
Neutral	0.722	0.716	0.700	0.708
Sadness	0.473	0.471	0.449	0.398
Surprise	0.560	0.577	0.539	0.373
Macro-F1	**0.520**	**0.369**	**0.434**	**0.421**
Weighted-F1	**0.628**	**0.529**	**0.606**	**0.564**
		EmoryNLP		
Joyful	0.538	0.524	0.532	0.464
Mad	0.391	0.438	0.408	0.400
Neutral	0.493	0.501	0.527	0.503
Peaceful	0.031	0.099	0.062	0.105
Powerful	0.110	0.114	0.097	0.048
Sad	0.362	0.392	0.335	0.332
Scared	0.264	0.403	0.320	0.062
Macro-F1	**0.274**	**0.309**	**0.228**	**0.239**
Weighted-F1	**0.355**	**0.388**	**0.372**	**0.315**
		DailyDialog		
Anger	0.420	0.479	0.359	0.430
Disgust	0.226	0.261	0.284	0.201
Fear	0.158	0.129	0.121	0.117
Happiness	0.538	0.560	0.551	0.455
Neutral	0.807	0.829	0.778	0.685
Sadness	0.249	0.244	0.272	0.146
Surprise	0.375	0.367	0.306	0.225
Macro-F1	**0.396**	**0.410**	**0.455**	**0.323**
Weighted-F1	**0.746**	**0.769**	**0.724**	**0.633**

Table 14: Per-emotion F1 scores across all datasets and models.

A.11 Annotater Agreement Distribution

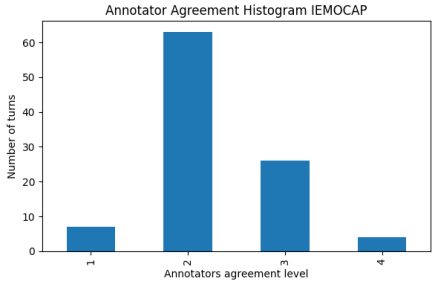


Figure 21: Annotator agreement histogram for IEMOCAP. The x-axis indicates the number of annotators (out of four) assigning the same emotion label to a turn, and the y-axis shows the number of turns.

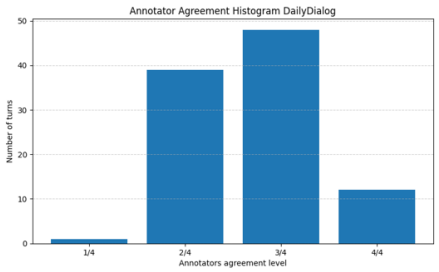


Figure 22: Annotator agreement histogram for DailyDialog. The x-axis indicates the number of annotators (out of four) assigning the same emotion label to a turn, and the y-axis shows the number of turns.

Figures 22 and 21 present the distribution of annotator agreement levels for DailyDialog and IEMOCAP, respectively. In both datasets, agreement is concentrated at intermediate levels rather than at full consensus. For DailyDialog, most turns fall into the 2/4 and 3/4 agreement categories, with very few instances of complete disagreement (1/4) or full agreement (4/4). This pattern reflects the prevalence of subtle, context-dependent utterances where multiple emotional interpretations remain plausible.

A similar trend is observed for IEMOCAP, where the majority of turns receive agreement levels of 2/4 or 3/4. Fully agreed-upon cases (4/4) are relatively rare, while a non-negligible number of turns exhibit low agreement (1/4), indicating substantial interpretative variability. Notably, minority emotion classes are over-represented in the lower agreement bins, whereas neutral and positive emotions are more likely to reach higher agreement levels.

Together, these distributions confirm that annotation ambiguity is a systematic property of ERC benchmarks rather than an artifact of individual datasets. The dominance of intermediate agreement levels suggests that many conversational utterances support multiple emotion plausibilities, challenging the assumption of a single unambiguous gold label.

A.12 Models' versions:

Table 15 summarizes the Large Language Models evaluated in this work, together with their versions, parameter scales, quantization schemes, and Ollama implementations. This selection spans a broad range of model sizes and architectures, enabling controlled comparison between large, compact, and safety-aligned LLMs under identical inference conditions.

Model	Version	Size (B)	Quantization	Ollama Tag
LLaMA-3	v3	70	Q4_O	`llama3:70b`
GPT-OSS	v1	120	MXFP4	`gpt-oss:120b`
Qwen	2.5	32	Q4_K_M	`qwen2.5:32b`
Mistral	v0.3	7	Q4_K_M	`mistral:7b`
Gemma	v3	27	Q4_K_M	`gemma:27b`
DeepSeek-R1	v1	70	Q4_K_M	`deepseek-r1:70b`
LLaMA-3	v3	8	Q4_K_M	`llama3:8b`
GPT-OSS-Safe	v1 (Safe)	120	MXFP4	`gpt-oss-safe:120b`

Table 15: Large Language Models used in this study, including model versions, parameter sizes, quantization schemes, and corresponding Ollama implementations.

B Linguistic Cues and Agreement Patterns

Figure 23 provides a qualitative breakdown of linguistic and contextual properties associated with high annotator agreement. These distributions support the observation that utterances with clearer intent, stronger contextual grounding, and explicit lexical or punctuation cues tend to yield more consistent emotion annotations.

LLM-as-Judge Confusion Matrices

Figure 24, Figure 25, and Figure 26 present the confusion matrices obtained for the LLM-as-Judge evaluation on DailyDialog, IEMOCAP, and EmoryNLP. Each matrix reports true negatives (TN), false positives (FP), false negatives (FN), and true positives (TP), where a positive prediction indicates that the LLM-as-Judge considers a candidate emotion semantically compatible with the human annotations.

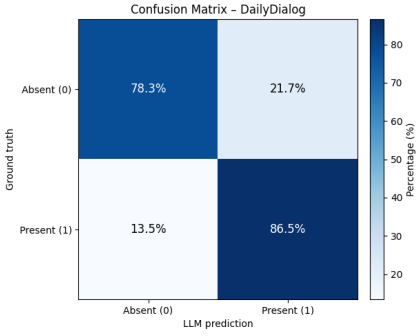


Figure 24: Confusion matrix for the LLM-as-Judge on DailyDialog.

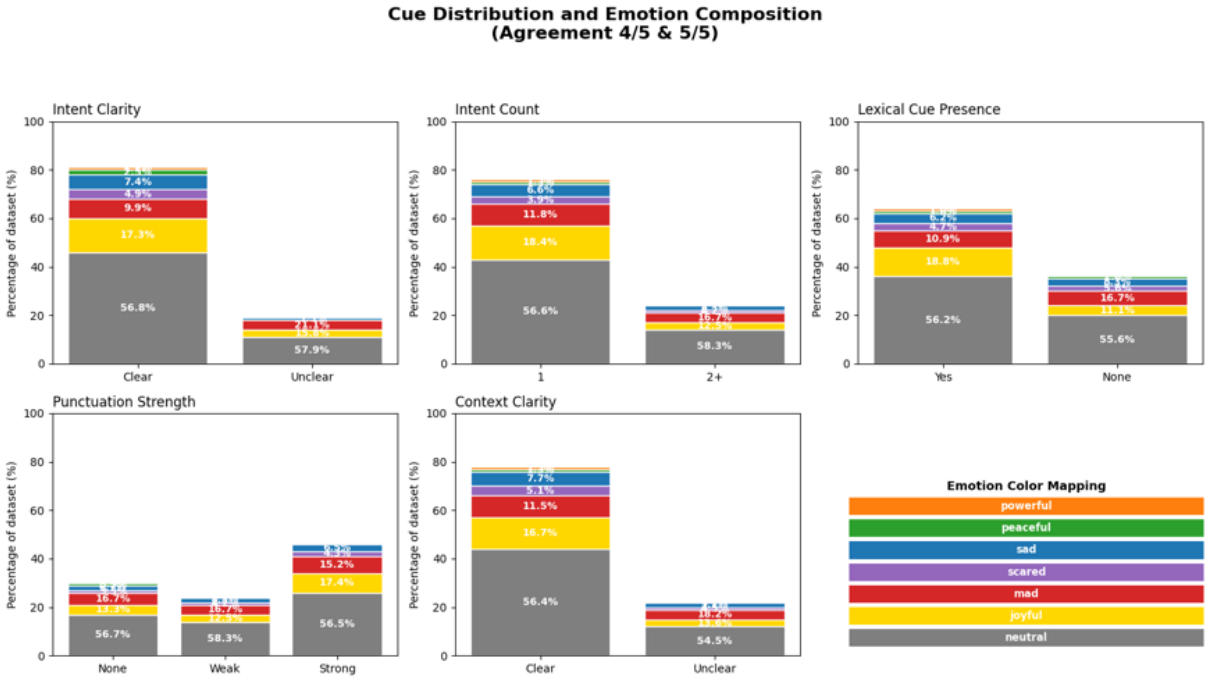


Figure 23: Distribution of linguistic and contextual cues for utterances from MELD with high annotator agreement (4/5 and 5/5). Each subplot shows the percentage of utterances exhibiting a given property (Intent Clarity, Intent Count, Lexical Cue Presence, Punctuation Strength, and Context Clarity), with stacked bars indicating emotion composition. The figure illustrates that high-agreement cases are predominantly associated with clear intent, single dominant intent, explicit lexical cues, and clear contextual grounding.

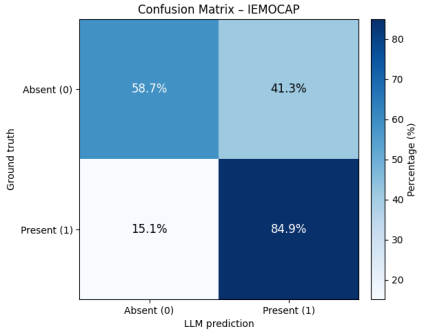


Figure 25: Confusion matrix for the LLM-as-Judge on IEMOCAP.

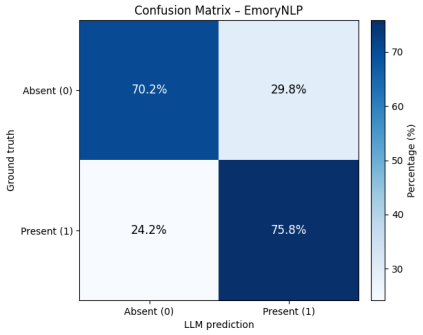


Figure 26: Confusion matrix for the LLM-as-Judge on EmoryNLP.

Across all three datasets, the confusion matrices reveal that the LLM-as-Judge achieves a strong balance between true positives and true negatives, confirming its ability to recognize emotions that are semantically compatible with human annotations. Notably, the dominant source of error is false positives, where the LLM-as-Judge accepts an emotion that was not selected by the majority of annotators. This behavior is particularly visible in IEMOCAP and EmoryNLP, where false positives are comparable in magnitude to true positives. Importantly, these false positives do not primarily reflect arbitrary mistakes: as shown in the main analysis, they are strongly concentrated in utterances with low inter-annotator agreement, indicating the presence of multiple plausible emotional interpretations rather than clear annotation errors.

C LLM-as-Judge Prompting Strategy (Binary Plausibility per Emotion)

Component	Description
System Instruction	You are an expert in emotion recognition in conversations. Your task is to determine whether a given emotion label is plausible for a target utterance, given the dialogue context.
Input Context	The full dialogue history preceding the target utterance, including speaker turns and conversational context.
Target Utterance	The specific dialogue turn whose emotional plausibility is being evaluated.
Candidate Emotion	One emotion label from the dataset label set (evaluated independently from other labels).
Task Instruction	Given the context and the target utterance, answer whether the candidate emotion is plausible. Respond with `Yes` or `No` only.
Output Format	Binary decision (`Yes`/`No`).
Decision Aggregation	An emotion is considered plausible for an utterance if the model answers `Yes` when queried with that emotion label.
Decoding Setup	Temperature = 0.3

Table 16: Prompting strategy used in the LLM-as-Judge framework, where emotion plausibility is assessed independently for each candidate label.