Exposing Weaknesses of Large Language Models in Emotion Recognition in Conversations

Anonymous ACL submission

Abstract

Emotion Recognition in Conversations (ERC) is essential for enabling machines to interpret human affect in dialogue. Existing benchmarks such as MELD, EmoryNLP, and DailyDialog each carry distinctive emotions and characteristics, which often bias models toward dataset-specific cues under fine-tuning. While Large Language Models (LLMs) report strong in-domain performance, it remains unclear whether they can recognize emotions in zero-shot scenarios without task-specific adaptation and when provided with full conversational context, including all preceding turns. We present the first systematic zero-shot evaluation of four LLMs (LLaMA-70B, Qwen-32B, Mistral-7B, GPT-OSS-120B) across the three benchmarks. Results show that aggregate metrics obscure consistent blind spots: minority emotions, which are systematically overlooked due to severe class imbalance, as well as short replies, interjections, negation, and sentence-type bias where exclamations and interrogations are prone to systematic misclassification, all of which are frequently misclassified. We provide a taxonomy of six recurring errors across models and datasets, revealing structural limitations in how LLMs process emotional nuance. Our findings underscore the need for error-sensitive evaluation and lay the groundwork for more robust ERC systems.

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), and is crucial for machines to understand dynamic human emotions

Dialogue:

Rachel: Guess what? I finally got the job I wanted!
Monica: Wow! That's amazing news.
Rachel: **Woohoo! I can't believe it actually happened!**

ERC Task:

Predict the emotion of Rachel's final utterance.
Labels: *Anger, Disgust, Fear, Joy, Neutral, Sadness, Surprise.*

True: Joy
LLaMA3-70B: Anger
Qwen-32B: Sadness
Mistral-7B: Neutral
GPT-OSS-120B: Joy

Interpretation:

Example of a joyful **interjection** from MELD style data. Despite the clear excitement signaled by "Woohoo!", three LLMs misinterpret it (Anger, Sadness, Neutral) while GPT-OSS-120B correctly predicts Joy. This highlights a common blind spot in emotion recognition.

Figure 1: Example of Intejection bias in **MELD**.

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, and DailyDialog (Poria et al., 2019; Zahiri and Choi, 2018; Li et al., 2017). 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 in-

formation 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). For example, MELD includes scripted sitcom dialogues enriched with persona information, where emotions are often exaggerated, while DailyDialog contains everyday conversations that lack such contextual details and instead convey more subtle affective cues (Li et al., 2017). Moreover, even though emotion recognition is framed as a classification task, each dataset defines its own distinctive set of emotions, which hinders cross-dataset transfer (Yang et al., 2022; Gao et al., 2022). As a result, models fine-tuned on one benchmark often adapt to dataset-specific biases rather than achieving genuine emotion understanding, consistent with findings that fine-tuning improves in-domain performance but weakens cross-domain generalization (Kang and Cho, 2024; Hu et al., 2023; Yang et al., 2022).

This raises the central question of our study: can LLMs recognize emotions in conversations across datasets in zero-shot settings, without task-specific training and including all preceding turns? Zero-shot evaluation probes intrinsic ability, uncovers weaknesses hidden by fine-tuning (Qin et al., 2023; Shen et al., 2025). Building on this motivation, our work provides a systematic investigation into how LLMs handle ERC in practice. The main contributions of this study are summarized as follows:

- The first systematic zero-shot evaluation of multiple LLMs across MELD, EmoryNLP, and DailyDialog, including all context.
- An automatic error analysis that identifies recurring patterns of misclassification, such as those hidden by overall metrics like accuracy or weighted F1.
- A typology of recurring failure modes across datasets, including minority emotions, interjections, negation, sentence-type bias (exclamation and interrogation), and utterance length.

2 Related Work

Emotion Recognition in Conversations has evolved from feature-based methods to neural and, more recently, LLM-based approaches. Early models used RNNs and graph networks to capture temporal and inter-speaker dynamics (Poria et al., 2017; Hazarika et al., 2018), while later works incorporated commonsense and psychological knowledge for better contextual reasoning (Zhong et al., 2019; Li et al., 2021). However, ERC benchmarks remain imbalanced, with neutral emotions dominating and minority emotions underrepresented (Poria et al., 2019; Yang et al., 2022). Recent LLM-based models like CoE (Shen et al., 2025) utilize scene and persona cues to enhance accuracy, though these improvements do not generalize well across datasets (Fu et al., 2024). The emergence of large language models such as LLaMA-30B (Touvron et al., 2023), Mistral-7B (Jiang et al., 2023), Qwen-32B (Bai et al., 2023), and GPT-OSS-120B (Patil et al., 2025) has reshaped ERC research. Previous studies have fine-tuned LLMs on specific ERC datasets (Zhang et al., 2023), achieving strong results but relying heavily on task-specific annotations. Zero-shot evaluation remains less explored (Winata et al., 2021). Our work builds on these efforts with a comprehensive zero-shot comparison across three ERC benchmarks (MELD, EmoryNLP, DailyDialog), uncovering consistent failure modes and highlighting the limitations of LLMs in understanding conversational emotions.

3 Experimental Setup

3.1 Datasets

We evaluate on three widely used ERC benchmarks: MELD, EmoryNLP, and DailyDialog. All are annotated for emotion classification but differ in genre, label sets, and class balance. MELD (Poria et al., 2019), derived from the *Friends* TV series, includes seven emotions (*anger, disgust, fear, joy, neutral, sadness, surprise*) but is heavily skewed toward *neutral* (>48%). EmoryNLP (Zahiri and Choi, 2018), also from *Friends*, uses a finer-grained scheme (*joyful, mad, peaceful, neutral, sad, powerful, scared*) with more subtle, context-dependent emotions. DailyDialog (Li et al., 2017) contains two-speaker everyday conversations annotated with seven emotions (*anger, disgust, fear, happiness, sadness, surprise, neutral*), However, over 80% of the utterances are labeled as neutral, reflecting a strong class imbalance. A comprehensive overview of dataset statistics can be found 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 = 0.0) to ensure consistency and comparability across models. Further details appear in Appendix A.3.

3.3 Evaluation Metrics

We report two complementary metrics. **Weighted F1 (W-F1)** reflects overall performance on imbalanced datasets but may 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 structural and methodological gaps. Most datasets are specific or scripted (e.g., MELD and EmoryNLP from *Friends*, DailyDialog from crowd-sourced text), limiting ecological validity and encouraging models to exploit stylistic cues rather than develop generalizable affective reasoning. Benchmarking practices further rely heavily on aggregate metrics (e.g. Accuracy or Weighted-F1) in many recent works, for example, CFN-ESA (Li et al., 2023a), PFA-ERC (Khule et al., 2024), and InstructERC (Lei et al., 2023) report only overall scores without in-depth analysis of minority-class performance. These metrics tend to obscure systematic errors on rare emotions such as *fear*, *disgust*, or *surprise*. Finally, most prior work evaluates a single model on a single dataset, making it difficult to assess cross-model robustness or dataset transferability. We address these gaps by evaluating multiple LLMs across three benchmarks in a zero-shot setting and analyzing recurring weaknesses shared across models and datasets.

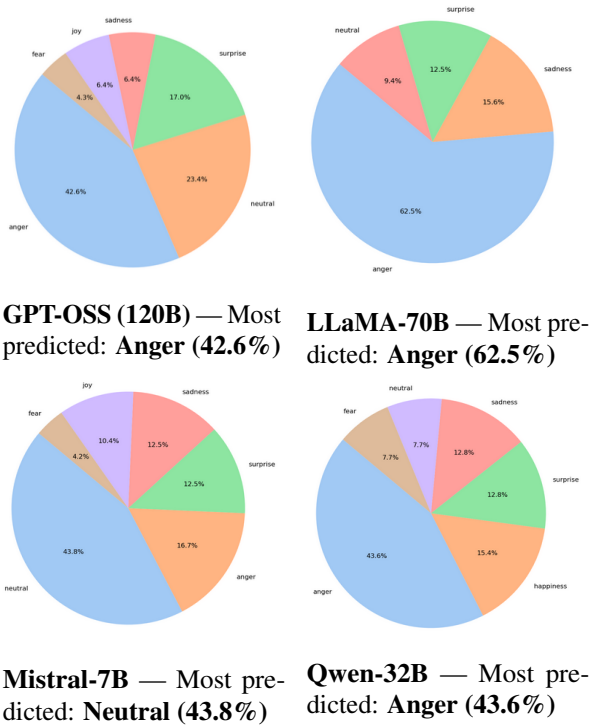
5 Results and Qualitative Analysis

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

5.1 Minority Emotions Analysis

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*.



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.

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 5). 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 10). Feature-based analysis in Appendix A.6 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

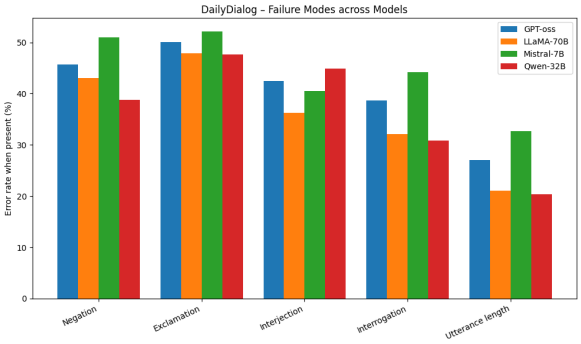


Figure 2: Comparison of failure-mode error rates in DailyDialog across four LLMs (GPT-OSS-120B, LLaMA-70B, Mistral-7B, Qwen-32B). Negations and exclamations show the highest error rates, followed by interjections and interrogations; fewer errors for short replies.

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 predictions, 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. 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.

6 Discussion

Our analysis reveals systematic error patterns across all models in Emotion Recognition in Conversations. Polarity inversions, misclassifications of interjections, exclamations and short replies are prevalent. These errors are largely caused by the models' reliance on surface-level cues such as punctuation, negation words, and very short utterances rather than deeper contextual understanding. This is quantitatively supported in Appendix A.6, where Figure 13 shows the distribution of linguistic features, such as negations across DailyDialog, EmoryNLP, and MELD. Additionally, minority emotions like Fear and Disgust are consistently misclassified as frequent emotions (e.g., Neutral or Anger), as shown in Table 10 and Appendix A.5, highlighting the bias toward dominant classes. Errors were automatically calculated using the feature-based analysis protocol (Appendix A.6), which detects linguistic features in utterances and computes conditional error rates. These findings confirm that current models struggle with emotional nuance and underscore the need for more robust, error-sensitive evaluation metrics in ERC.

7 Conclusion

This paper presents the first systematic zero-shot evaluation of LLMs on emotion recognition in conversations across MELD, EmoryNLP, and DailyDialog. Our analysis uncovers consistent blind spots, including short utterances, interjections, negations, and sentence-type biases, which repeatedly lead to misclassifications and hinder the accurate capture of emotional nuance. By combining quantitative metrics with feature-based analyses, we demonstrate that these errors are systematic and often driven by dataset characteristics. These findings underscore the need for error-sensitive evaluation protocols and motivate the development of more robust, contextually grounded benchmarks to better assess conversational emotion recognition.

8 Limitations

This study has several limitations. Our analysis is primarily diagnostic, identifying failure modes in LLMs without proposing specific solutions or fixes. We evaluate only three text-based benchmarks (MELD, EmoryNLP, and DailyDialog), excluding multimodal datasets like IEMOCAP, which could provide richer insights by incorporating nonverbal cues such as tone of voice, facial expressions, or gestures. Additionally, we do not assess how biases introduced during pretraining, or exposure to scripted data, might influence model behavior and result in misclassifications. Future work should explore the inclusion of multimodal cues to enhance emotional understanding and investigate targeted mitigation strategies to address the identified blind spots, such as refining model architectures or introducing domain-specific training techniques to reduce reliance on surface cues.

References

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

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*.

Zhen Gao, Ming Zhou, Rui Zhang, and Wenjie Wang. 2022. Emotion-shift aware contrastive learning for emotion recognition in conversations. In *EMNLP*.

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*.

Dou Hu, Yinan Bao, Lingwei Wei, Wei Zhou, and Songlin Hu. 2023. Supervised adversarial contrastive learning for emotion recognition in conversations. In *ACL*.

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

Yujin Kang and Yoon-Sik Cho. 2024. Improving contrastive learning in emotion recognition in conversation via data augmentation and decoupled neutral emotion. In *EACL*.

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*.

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 2: 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 three widely used Emotion Recognition in Conversation (ERC) benchmarks: DailyDialog, MELD, and EmoryNLP. The figures below present the distribution of emotions within each dataset. 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.

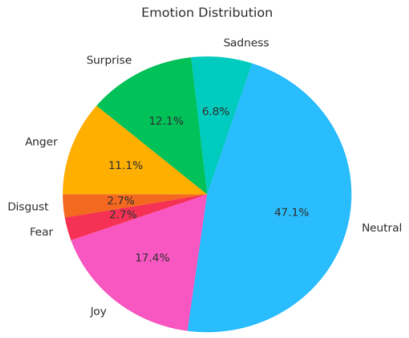


Figure 3: Emotion distributions in MELD,

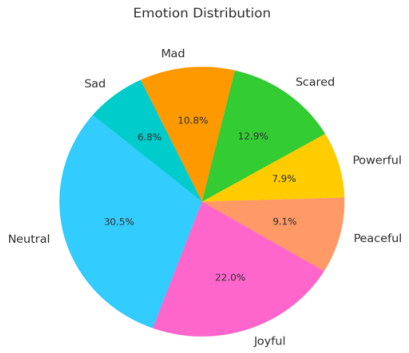


Figure 4: Emotion distributions in EmoryNLP.

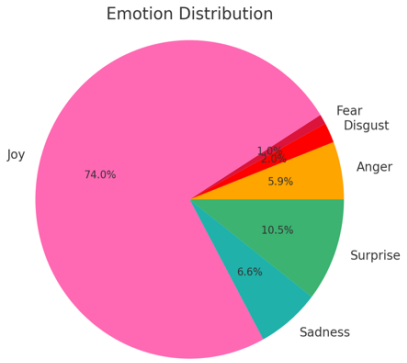


Figure 5: Emotion distributions in DailyDialog.

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 3. 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 3: 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 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.

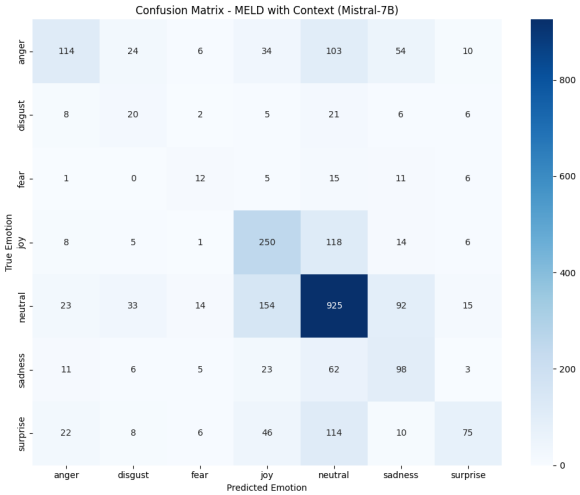


Figure 6: **MELD – GPT-OSS-120B. Fear** and **Disgust** are almost never recognized correctly and are largely predicted as **Neutral** or **Anger**, indicating strong attraction to dominant emotion categories. **Joy** and **Surprise** also show overlap, reflecting surface-form confusion from exclamatory expressions. The strong diagonal for **Neutral** inflates weighted-F1 despite weak coverage of minority classes.

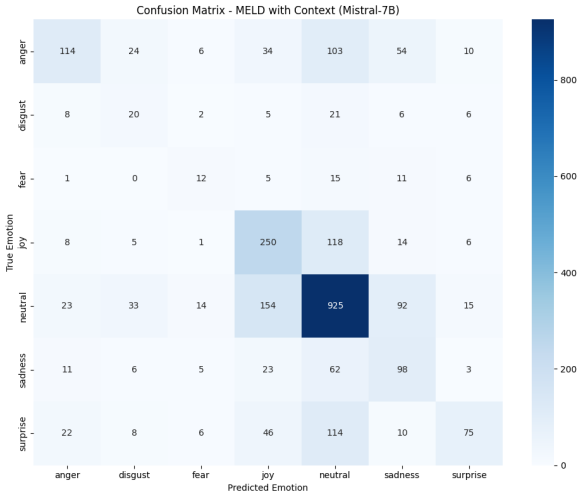


Figure 7: **MELD – Mistral-7B.** Minority emotions such as **Fear** and **Disgust** collapse almost entirely into **Neutral**, while **Sadness** is frequently misread as **Joy**. This reveals a polarity-bias effect: low-valence emotions are absorbed into higher-valence or neutral predictions. Although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

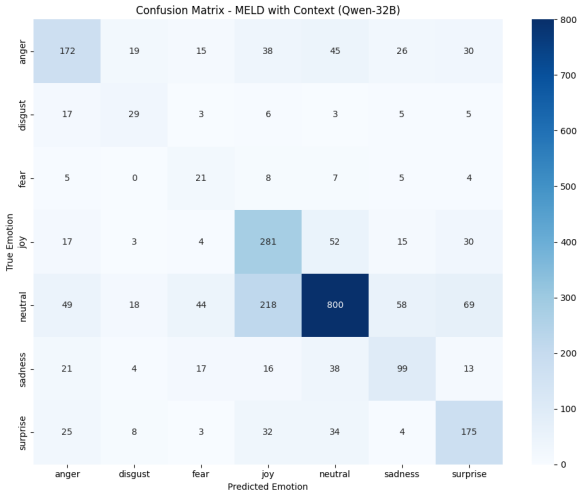


Figure 8: **MELD – Qwen-32B.** While **Joy** and **Neutral** are more distinct, rare classes (**Fear**, **Disgust**) continue to be absorbed by **Anger** or **Neutral**. **Sadness** occasionally shifts toward **Joy**, suggesting confusion in valence polarity rather than arousal intensity. These errors reinforce that higher model scale does not guarantee balanced emotional discrimination.

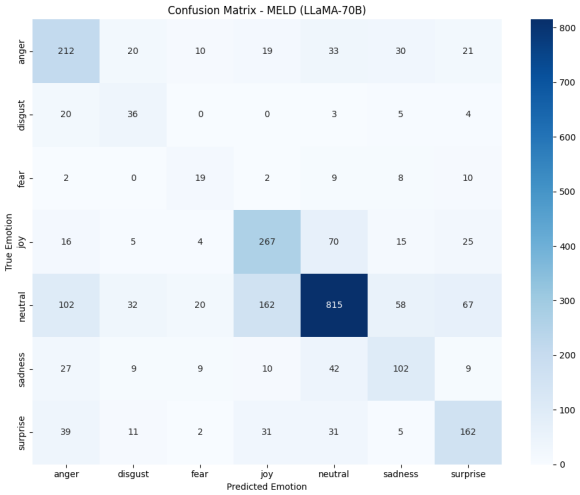


Figure 9: **MELD – LLaMA-70B.** LLaMA exhibits the clearest **Neutral** diagonal and relatively stable **Joy–Surprise** boundary, but still confuses **Disgust** and **Fear** with **Anger**. The persistence of this confusion across all models highlights a structural limitation in emotion separation—particularly for low-frequency or context-dependent states.

A.6 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.

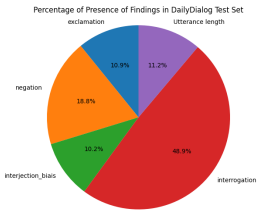


Figure 10: DailyDialog

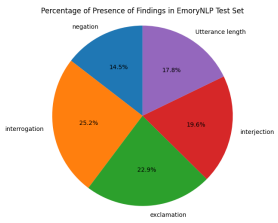


Figure 11: EmoryNLP

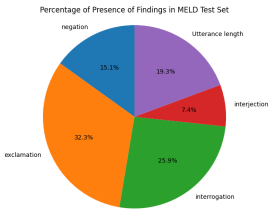


Figure 12: MELD

Figure 13: 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.

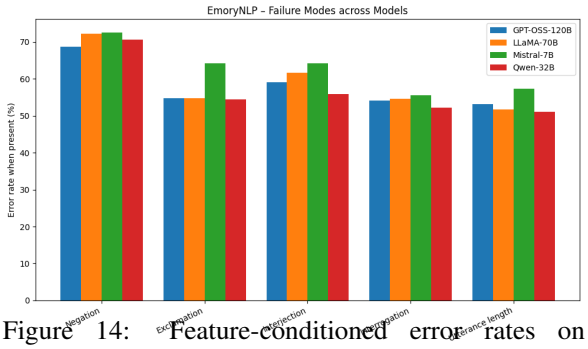


Figure 14: Feature-conditioned error rates on EmoryNLP across GPT-OSS-120B, LLaMA-70B, Mistral-7B, and Qwen-32B.

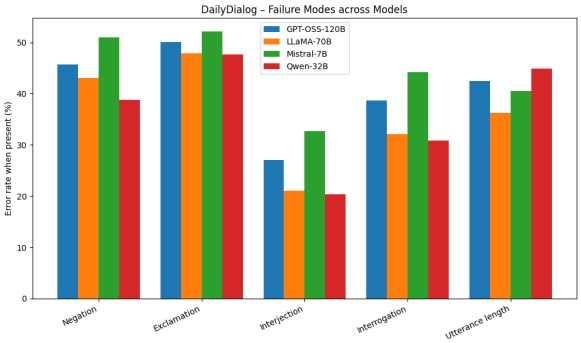


Figure 15: Feature-conditioned error rates on DailyDialog across GPT-OSS-120B, LLaMA-70B, Mistral-7B, and Qwen-32B.

A.7 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 4: 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 5: 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 6: 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 7: 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**

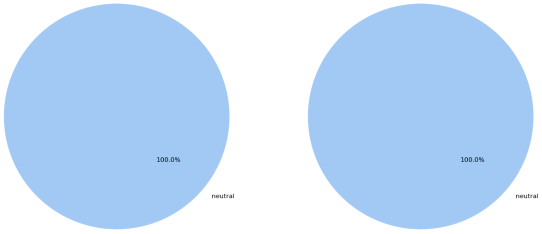
Table 8: Example from **DailyDialog** using **GPT-OSS-120B**. Type of error: *Short reply ambiguity* — minimal responses ("Yeah", "Nothing special") are interpreted inconsistently, oscillating between **joy** and **sadness**. This inconsistency indicates weak sensitivity to discourse pragmatics.

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 9: 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.

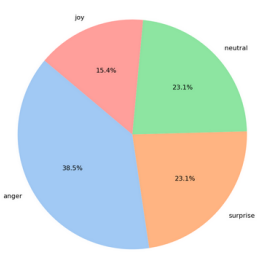
A.8 Emotion–Feature Interaction Analysis

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.

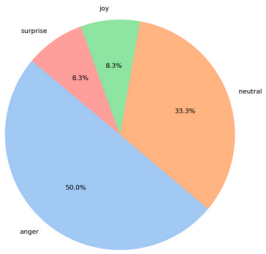


(a) LLaMA-70B **(b)** GPT-OSS-120B

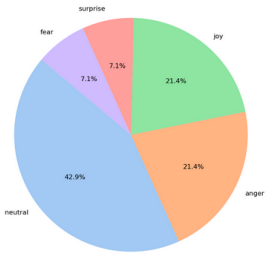
Figure 16: Distribution of predictions for **Fear** utterances containing interjections in DailyDialog. Both **LLaMA-70B (a)** and **GPT-OSS-120B (b)** completely misclassify these instances as **Neutral (100%)**, showing a collapse of recognition for this emotion when interjections appear.



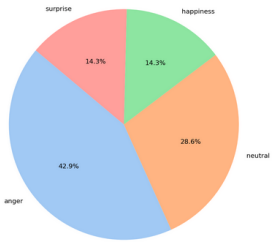
GPT-OSS (120B) — Most predicted: **Anger (38.5%)**



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

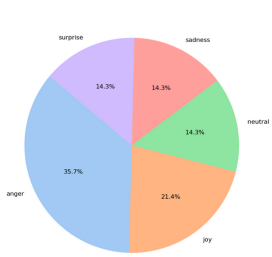


Mistral-7B — Most predicted: **Neutral (42.9%)**

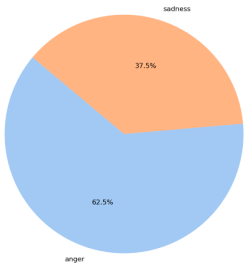


Qwen-32B — Most predicted: **Anger (42.9%)**

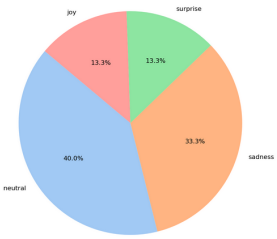
Figure 17: Distribution of predictions for **Sadness** utterances containing exclamations across all four models. While the ground truth emotion is **Sadness**, all models show strong confusion toward high-arousal categories such as **Anger** and **Neutral**. This suggests that the presence of exclamations biases models toward interpreting the emotional tone as more intense or externally directed, leading to systematic misclassification of low-arousal emotions like sadness.



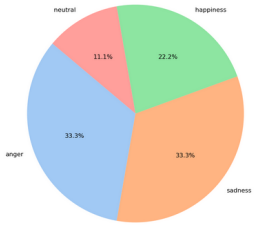
GPT-OSS (120B) — Most predicted: **Anger (35.7%)**



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



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



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

Figure 18: 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.9 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 10: Per-emotion F1 scores across all datasets and models.