Real-Time Mitigation of Negative Emotion in Customer Care Calls

Anonymous ACL submission

Abstract

Speech emotion recognition (SER) is a compelling yet challenging research area with substantial practical relevance, particularly in enhancing human–machine interaction. Despite considerable progress in the field, the scarcity of realistic datasets that reflect real-world conditions makes it difficult to analyze system behavior in practice and can lead to degraded performance in industrial applications. In this study, we propose a system that detects negative emotions at each turn in a conversation by leveraging both linguistic and acoustic features. The approach is evaluated on real-world data, with a particular focus on identifying and responding to negative emotion in customer support scenarios. Designed for real-time application, the system is suitable for live deployment in call center environments. Furthermore, we propose an effective prompting strategy for using large language models (LLMs) as annotators, generating labeled data used to fine-tune small language models that achieve performance on par with the LLM used for annotation, while remaining suitable for real-time deployment.

1 Introduction

Automatic emotion recognition using modalities such as text, speech, image, and video has been an active research area for several decades, with significant advancements achieved through the application of increasingly sophisticated models. Recognizing emotions from speech is particularly valuable in diverse human–machine interactions, including conversational agents, virtual assistants, and interactive systems. Although numerous studies report promising evaluation outcomes on public benchmark datasets (Caralt et al., 2025; Shou et al., 2024; Xue et al., 2024; Fu, 2024; Lei et al., 2023; Stricker and Paroubek, 2024; Ma et al., 2023; Lin et al., 2023), the field continues to face a shortage of realistic data and practical methods suitable for real-world deployment.

This study investigates speech emotion recognition (SER) in spoken dialogue systems, with a particular focus on enterprise customer support. In this domain, users interact with automated or virtual agents to resolve service or product-related issues. Accurately detecting emotions in such interactions can enhance user experience, support timely intervention, and improve the overall effectiveness of automated support systems. The principal aim is to identify and appropriately respond to negative emotional states like anger or frustration by employing mitigation strategies, such as escalating the call to a human agent or incorporating more empathetic system responses.

A critical aspect of this work is the requirement for real-time processing, which significantly increases the complexity of the task. In this setting, "real-time" refers to the ability to detect emotions with low CPU latency as the conversation progresses, rather than after it concludes. This imposes strict constraints on system responsiveness, as predictions must be made quickly enough to support timely interventions during the interaction. Consequently, the deployment of highly complex models becomes impractical, as they often introduce delays that are incompatible with the demands of live dialogue systems. The rapid advancement and application of large language models (LLMs) have significantly influenced a wide range of research domains, including SER. While LLMs offer substantial benefits for classification tasks particularly in low-resource scenarios where limited or no labeled data is available their use is associated with considerable computational cost and latency. These limitations pose significant challenges for practical, large-scale SER deployment. Therefore, in this work, we leverage the capabilities of LLMs in an offline setting to generate and curate training data for SER, thereby avoiding the latency and cost constraints associated with real-time inference.

A core challenge in speech emotion recogni-

tion (SER) is the scarcity of comprehensive, high-quality datasets. Most publicly available corpora are limited in size and primarily consist of acted speech recorded in controlled settings, which lack the natural variability and spontaneity characteristic of real-world emotional expression. Furthermore, imbalanced emotion classes and insufficient demographic diversity impede the robustness and generalizability of trained models (Garg et al., 2024; Upadhyay et al., 2025; Lin et al., 2024). Although real-world, in-domain data is often available in industry, for example, through customer service interactions or voice-enabled platforms, such data are typically unlabeled and proprietary, and annotation at scale is both costly and time-consuming. To address this limitation, we leverage LLMs to generate labeled training data by guiding them with an annotation protocol originally developed for human labeling of evaluation data.

To identify negative emotions such as caller frustration, we integrate both linguistic and acoustic cues, as each provides complementary information. Linguistic indicators include semantic signals such as negative sentiment or intent, as well as explicit expressions like angry language, complaints about the system, repeated requests, and swearing. Acoustic features capture prosodic and paralinguistic signals, including elevated pitch, increased volume, vocal strain, abrupt changes in loudness, and variations in speaking rate. By combining these modalities, the system is better equipped to detect subtle and context-dependent expressions of negative affect. The contributions of this paper are summarized as follows: (1) We develop a low CPU latency, efficient, production-ready solution for real-time detection of negative customer emotions, integrating both linguistic and acoustic cues in speech; (2) We leverage prompt tuning to utilize large language models (LLMs) as offline annotators for generating labeled training data, which is then used to fine-tune lightweight LLMs suitable for low-latency deployment; (3) We conduct a comparative evaluation of multiple modeling approaches on both publicly available datasets and real-world industry data to assess performance under different conditions.

2 Related Work

Recent advances in large language models (LLMs) have shifted emotion recognition from traditional discriminative models to generative paradigms across modalities (Ma et al., 2025; Zhang et al., 2024; Broekens et al., 2023). Techniques such as chain-of-thought reasoning, contextual modeling, in-context learning, and fine-tuning have further improved automatic emotion recognition performance (Bhaumik and Strzalkowski, 2024; Zhang et al., 2023; Peng et al., 2024). While some studies explore multi-modal fusion of acoustic and textual features (Dutta and Ganapathy, 2025; Wu et al., 2025), most work remains text-centric. A growing body of research applies LLMs to speech emotion recognition (SER) (Xu et al., 2024; Gong et al., 2023; Chen et al., 2025), often relying on speaker identity, context, or auxiliary tasks like speaker prediction (Lei et al., 2023; Xue et al., 2024), which limits their suitability for real-time applications with unknown participants and limited context.

In task-oriented dialogues, emotion recognition typically prioritizes detecting negative states—such as frustration, anger, or dissatisfaction—rather than covering a broad emotional spectrum. Prior work has used LLMs with in-context learning to identify frustration from transcripts (Caralt et al., 2025), or relied on conversational patterns and agent behaviors to infer dissatisfaction. In this work, we focus on real-time SER for task-oriented conversational systems in commercial settings. Our goal is to detect negative emotional states using only the audio of the current customer utterance at inference time. To ensure practical deployment, we avoid reliance on dialogue history, speaker identity, or other contextual metadata that may be unavailable or introduce system complexity. The following sections detail our proposed method, dataset construction process, results, and discussion.

3 Methodology

We propose a low-cost, low-latency SER approach that is developed using spontaneous, in-domain data from real customer care calls. Large language models are employed as annotators to label production data, which is then used to fine-tune small language models (SLMs) for efficient inference.

3.1 Leveraging Large Language Models for Offline Data Annotation

Large language models (LLMs) have shown strong performance in various natural language and speech tasks, including SER. A key challenge in SER is the limited availability of high-quality, in-domain labeled data for training robust models.

Traditional annotation methods, which rely on manual labeling, are time-consuming and costly, particularly for large datasets needed to capture the variability in natural speech.

To address this challenge, we adopt an offline approach in which a large language model, more precisely Gemini 2.0 Flash, is employed to automatically annotate training data for emotion recognition. The resulting annotated dataset is then used to fine-tune SLMs designed for efficient real-time prediction. This strategy enables us to benefit from the language understanding capabilities of LLMs while avoiding their high inference latency and computational requirements during deployment. We compare the performance of different zero-shot prompting approaches to improve the performance of the model as annotator. These strategies are optimized over the ProdEval test set discussed in Section 4.1. Table 1 shows the different prompting approaches that we used for our experiments. In the Simple approach, we prompt the model with an annotation guide that was originally developed for human annotators to ensure consistency and alignment between machine-generated and human-generated labels. However, the previous context of the utterance is not provided. In ZCOT_Simple, we use the zeroshot chain-of-thought (COT) approach (Kojima et al., 2022) where the model is first supposed to generate a reasoning before arriving at the result, thus improving the prediction performance of the model. In ZCOT_Context, along with the current customer utterance, we provide a context window that contains previous agent-customer interactions that are not available at inference time, while the rest of the prompt remains the same as ZCOT_Simple. In ZCOT_Detailed_Context, the context window consists of utterance sequence numbers, and dialog state ids (DSIDs) along with the previous agent-user interactions.

In PAS_Detailed_Context, we use the Plan and Solve prompting strategy (Wang et al., 2023), which improves the zeroshot COT reasoning, by specifically prompting the model to first understand the problem, then plan a strategy to solve the problem and finally follow the strategy to solve the problem. Finally, in RP_Detailed_Context, we use the roleplay prompting strategy (Kong et al., 2024), wherein the model is prompted to act as an intent analyst following which the model will play the role of an intent analyst to recognize the emotion from the utterance. In both the prior and latter approaches, the context window is same as ZCOT_Detailed_Context approach. The prompt for RP_Detailed_Context is shown in Figure 1. While, the prompts for other approaches are shown in Appendix A.

> From now on, you are an intent analyst tasked with analyzing audio from callers to detect their emotional state. Your analysis should categorize the caller's emotion as either Non-Negative or Negative. Your analysis till today has helped to resolve many caller issues. You actively monitor for indications of negative emotion throughout the call, including at the beginning of the interaction. You use the below definitions of negative emotion and neutral emotion to determine the emotion.
>
> Negative Emotion is expressed through anger, irritation, dissatisfaction, etc. The cues that show negative emotion could be linguistic, acoustic, or both. Linguistic cues include negative/angry words, swearing, complaining about the system or its suggestions, repeating the same request multiple times, etc. Acoustic cues include changes in the tone of the speaker or other voice changes including: a higher pitch, louder volume, more tension or strain which can sound tight or harsh, or sudden changes in loudness or speed. In case of acoustic cues pay attention to cases where the pitch of the caller's voice is not calm.
>
> Non-Negative Emotion if the caller does not seem to have a negative emotion in that utterance.
>
> You are doing a binary labeling here, so you mark the positive emotions as Non-Negative Emotion. You should provide a concise determination.

Figure 1: Prompt used for RP_Detailed_Context prompting approach

We empirically determined that annotation performance is optimal when the context window consists of a maximum of three previous agent-customer interactions along with the current customer utterance. The results of the annotation performance of Gemini 2.0 Flash using Simple approach for different context window sizes is shared in Appendix B.

3.2 Fine-tuning SLMs

Running a foundational model for real-time SER is not feasible, as it will involve high cost and latency. Instead, we need to focus on small language models (SLMS) that are finetuned only for SER. To this end, we use the train dataset annotated using the foundational model to finetune a smaller audio-based language model, more specifically, Qwen 2 Audio 7B Instruct (Chu et al., 2024). We use LoRA (Hu et al.) to finetune the model. Since, our existing legacy system is built to process only the current user's utterance in real time, so we use only the audio of the customer utterance as input to the model and do not add any other context to

Approach Name	Prompting Strategy	Context	Input	Input format
Simple	Simple prompt	None	Agent prompt + Current utterance	Audio
ZCOT_Simple	Zero-shot COT			
ZCOT_Context		Last 3 agent-customer interactions	Context + Agent prompt + Current utterance	Text + Audio
ZCOT_Detailed_Context		Last 3 agent-customer interactions, DSIDs and utterance sequence no.		
PAS_Detailed_Context	Plan and Solve			
RP_Detailed_Context	Roleplay			

Table 1: Summary of prompting approaches for using Gemini 2.0 Flash for annotation

the input.

3.3 Reduction of Inference Time CPU Latency

SLMs having 7B paramters have low inference time latency on GPU. However, on CPU the inference time latency increases considerably. Although model compression methods such as bitsandbytes (Dettmers et al., 2021) are available, they are optimized to reduce the inference time latency in GPUs. Since our solution focuses on reducing CPU inference time latency, we need to use SLMs that are close to 1B-2B parameters. Thus, we use the knowledge distillation method using the fine-tuned Qwen 2 Audio 7B Instruct as the teacher model and Aero 1 Audio [1], which is an audio-based SLM with 1.5B parameters as the student model. Since the vocabulary size of both the teacher and student models is different, we use universal logit distillation (ULD) (Boizard et al.) which calculates the Wasserstein distance to minimize the distance between the soft probabilities of the teacher and student models.

4 Datasets

Our goal in this study is to present evaluation results that accurately reflect the production environment and real-world data. To this end, we report SER accuracies, ROC-AUC, class-wise precision and recall on two internal datasets collected through distinct methodologies and provide an analysis of the observed differences. Additionally, we provide results on two widely used publicly available datasets for comparative analysis.

4.1 Internal Test Sets

A key challenge in SER is the strong class imbalance present in both training and real-world evaluation data, where emotionally neutral or low-arousal states dominate, and emotionally expressive states are relatively rare. This imbalance hinders model learning and reliable performance assessment, making it crucial to construct evaluation sets that fairly represent both classes, especially in production settings where detecting negative affect is a primary concern.

Throughout this paper, we refer to emotion classes as negative and non-negative, rather than the conventional neutral vs. emotional split. This reflects our focus on detecting negative states (e.g., anger, frustration), which are most relevant in customer support. In the customer care domain, other emotions such as happiness or surprise occur too infrequently to support consistent labeling and are grouped under the non-negative class.

ProdEval This is our most realistic evaluation set, constructed from real production conversations. As noted earlier, a key challenge in emotion data selection is the natural class imbalance. In our application, analysis of randomly sampled calls shows that fewer than 1% of conversations exhibit clear negative emotions. To ensure meaningful evaluation, we applied class balancing during the selection process to include sufficient samples of both negative and non-negative classes.

In order to make expert annotation more efficient, we utilized a data selection strategy that increased the proportion of negative emotion calls without increasing the overall number of calls to be labeled. To achieve this, we applied a rule-based filtering approach informed by domain knowledge. Specifically, we constructed queries over production data to identify calls likely to contain negative emotions, focusing on conversational patterns associated with system misunderstandings, one of the most common sources of caller frustration in our application. To reduce potential bias from targeted sampling, we supplemented these with randomly selected calls, resulting in a distribution with a higher proportion of negative emotion compared to the natural distribution observed in production data.

The selected evaluation data was annotated by

[1] https://huggingface.co/lmms-lab/Aero-1-Audio

domain experts familiar with the specific characteristics and user behaviors of the target application. Because the system is designed to detect negative emotion at the level of individual user turns, annotation was performed at the utterance level, enabling fine-grained supervision aligned with the model's inference granularity. To support accurate labeling, each utterance was presented along with the full audio of the call and all preceding turns, allowing annotators to consider both acoustic cues and conversational context when making judgments.

CueEval This evaluation set was also derived from production conversations, but it differs from ProdEval in both data selection and annotation methodology. As with ProdEval, random sampling was avoided due to the highly skewed distribution of emotional expressions in natural data. Instead, the selection strategy focused on leveraging existing production-side labels generated through a combination of human-in-the-loop processes and model-based inference that indicated instances of swearing or anger. These segments were combined with interactions labeled neutral or positive based on customer satisfaction cues, such as expressions of gratitude toward the system. The resulting subset was then re-annotated by expert raters in an offline setting to ensure consistent, high-quality emotion labels.

Although CueEval reflects real production utterances, it does not capture the full realism of ProdEval. This limitation stems from its selection criteria, which primarily targeted linguistic cues such as swearing or overt expressions of anger—based on existing production labels. In contrast, ProdEval includes more subtle and context-dependent expressions of negative emotion, such as very short utterances (e.g., "yes" or "no") that carry emotional charge but are challenging to detect automatically. These instances more accurately reflect the nuanced nature of real-world interactions. Furthermore, unlike ProdEval, where annotators had access to the full call audio and interaction history, CueEval annotations were performed at the utterance level without full conversational context, potentially limiting labeling accuracy for ambiguous cases.

4.2 Public Test Sets

We also evaluate our prediction models on two publicly available SER datasets: MELD (Poria et al., 2019) and IEMOCAP (Busso et al., 2008). We only perform evaluation on the test split and do not use the train split to train our models specifically to perform prediction on these datasets. MELD dataset consists of utterances from the Friends TV series, it consists of the following labels: anger, disgust, sadness, joy, neutral, surprise and fear. IEMOCAP dataset consists of dyadic sessions where the emotional utterances of an actor are based on hypothetical scenarios. It consists of the following labels: happiness, anger, sadness, frustration and neutral. Similar to Caralt et al. (2025), in MELD, we consider the emotion labels: anger and disgust as negative emotion and thus relabel these emotions as negative emotion while the rest of the emotion labels are relabeled as non-negative emotion. So also, in IEMOCAP, we consider the emotion labels: anger, disgust and frustration as negative emotion while the rest of the emotion labels are relabeled as non-negative emotion

4.3 Train Dataset

The train dataset is created using the same strategy that we used to create ProdEval. We use the role-play prompting strategy to annotate the utterances in the train dataset. After performing the annotation, we further reduced the class imbalance of negative and non-negative emotions by using 1:K negative sampling.

Table 2 shows the number of samples for negative and non-negative classes across all the datasets.

Datasets	Non-Negative	Negative
ProdEval	1030	566
CueEval	114	92
MELD	1256	413
IEMOCAP	551	384
Train Dataset	25593	6447

Table 2: Statistics of negative and non-negative labels in the test and train datasets

5 Results and Discussion

5.1 Experiment Settings

We use Gemini 2.0 Flash as our annotator, leveraging its multi-modal capabilities to enhance annotation quality. For model comparison, we evaluate our fine-tuned Qwen 2 Audio 7B Instruct (Qwen 2 Audio) and distilled Aero 1 Audio (Aero) models against both text- and audio-based baselines. In the text-based setting, we use BERT base uncased (BERT) (Devlin et al., 2019), Llama 3.1 8B Instruct (Llama 3.1) (Dubey et al., 2024), Qwen 2 7B In-

struct (Qwen 2 Text) (Team et al., 2024), and Qwen 3 4B Instruct (Qwen 3 Text) (Yang et al., 2025) as our baseline models. These models are trained on our dataset and evaluated via ASR-transcribed utterances. The text-based SLMs: Llama 3.1, Qwen 2 Text and Qwen 3 Text are finetuned using LoRA.

Among audio models, we include SER-Odyssey-Baseline-WavLM-Arousal (WavLM-Arousal) (Goncalves et al., 2024), classifying utterances with arousal scores >0.5 as negative. We also test Kimi Audio 7B Instruct (Kimi Audio) (Team, 2024)—a Qwen 2.5-based model via zero-shot inference due to LoRA incompatibility. Finally, we include Gemini 2.0 Flash (Gemini) as a zero-shot baseline using roleplay prompting without dialogue context. All audio models are evaluated using only the current customer utterance. Prompts used for text-based SLMs are detailed in Appendix C. Prompts used for Gemini and audio-based SLMs are detailed in Appendix D. Given the class imbalance in SER, we report precision and recall per class, along with ROC-AUC and accuracy for overall comparison. Recall is particularly important in our use case, as the goal is to detect negative customer emotions in customer care calls. Failing to identify angry or frustrated customers can result in unresolved issues or escalations, making high recall in negative emotion classes a critical priority.

5.2 Experimental Results

5.2.1 Performance of Gemini 2.0 Flash as annotator

The performance of Gemini 2.0 Flash as an annotator on ProdEval data is shown in Table 3. Note that when evaluating an LLM as an annotator, unlike during inference, where recall is often more important than precision, the goal is to achieve high performance on both precision and recall. This is because low precision will lead to mislabeled training samples. We observe that RP_Detailed_Context approach gives the highest accuracy and the ROC-AUC score is close to ZCOT_Detailed_Context approach. The recall of RP_Detailed_Context approach on negative emotion is also slightly lower than ZCOT_Detailed_Context approach. The precision is better than ZCOT_Detailed_Context approach and only lower than ZCOT_Simple approach. However, the ZCOT_Simple approach has the lowest recall on negative emotion. The performance

Approach Name	**ROC-AUC**	**Accuracy**
Simple	65.16	72.27
ZCOT_Simple	65.61	73.52
ZCOT_Context	69.57	72.33
ZCOT_Detailed_Context	69.99	73.27
PAS_Detailed_Context	67.24	72.96
RP_Detailed_Context	69.18	**74.34**
Negative Emotion		
Approach Name	**Precision**	**Recall**
Simple	68	41
ZCOT_Simple	75	38
ZCOT_Context	61	60
ZCOT_Detailed_Context	63	59
PAS_Detailed_Context	67	48
RP_Detailed_Context	68	51
Non-Negative Emotion		
Approach Name	**Precision**	**Recall**
Simple	73	90
ZCOT_Simple	73	93
ZCOT_Context	78	79
ZCOT_Detailed_Context	78	81
PAS_Detailed_Context	75	87
RP_Detailed_Context	76	87

Table 3: The performance of Gemini 2.0 Flash as an annotator when using different approaches to prompt the LLM.

of all the approaches on non-negative emotion is comparable. ZCOT_Simple, ZCOT_Context, and ZCOT_Detailed_Context use the zershot COT prompting, wherein the model first generates the reasoning based on the prompt, after which it uses the prompt and reasoning to generate the emotion, effectively being a two-pass approach (Kojima et al., 2022) and in the process also increasing the cost of output generation when used in scale. In contrast, RP_Detailed_Context uses the roleplay approach where assigning the role of intent analyst leads to implicit COT and also provides better reasoning than zero-shot COT (Kong et al., 2024), thus effectively being a single-pass approach. Thus, in case of data annotation at scale, RP_Detailed_Context approach is suitable as its ability to detect negative emotion is comparable to that of ZCOT_Detailed_Context while also giving highest overall accuracy and lowering the cost of output generation. Furthermore, in ProdEval the Cohen's Kappa score between predictions using RP_Detailed_Context and human annotator is 0.4 indicating fair agreement between Gemini as an annotator and human annotator. Thus, showing the effectiveness of our prompting method for large scale annotation. The prediction performance

	ROC-AUC				Avg. ROC-AUC	Accuracy				Avg. Accuracy
Models	**ProdEval**	**CueEval**	**MELD**	**IEMOCAP**		**ProdEval**	**CueEval**	**MELD**	**IEMOCAP**	
Text Models										
BERT	61.87	68.44	55.08	60.64	61.51	69.57	68.76	68.48	57.75	66.14
Llama 3.1	62.8	69.7	69.6	66.19	**67.07**	69.63	68.44	61.47	69.09	**67.16**
Qwen 2 Text	61.91	64.79	67.9	68.02	65.66	68.69	63.59	58.65	64.15	63.77
Qwen 3 Text	62.03	68.09	69.66	65.02	66.20	69.57	67.47	63.39	67.37	66.95
Audio Models										
WavLM-Arousal	70.47	53	58.7	77.51	64.92	75.25	54.36	41.4	72.19	60.80
Kimi Audio	71.29	83.78	73.52	79.52	**77.03**	76.28	84.95	81.3	76.14	**79.67**
Gemini	64.84	88.19	71.3	74.79	74.78	72.02	87.86	73.81	73.68	76.84
Qwen 2 Audio	66.42	81.49	65.92	69.95	70.95	72.2	80.09	57.27	72.4	70.49
Aero	67.83	77.68	62.12	69.54	69.29	73.21	76.69	47.51	72.62	67.51
Precision and Recall for Negative Emotion										
	Precision				**Avg. Precision**	**Recall**				**Avg. Recall**
Models	**ProdEval**	**CueEval**	**MELD**	**IEMOCAP**		**ProdEval**	**CueEval**	**MELD**	**IEMOCAP**	
Text Models										
BERT	63	63	34	73	58.25	35	72	29	44	45
Llama 3.1	61	80	38	70	**62.25**	39	58	86	82	66.25
Qwen 2 Text	59	57	36	68	55	39	76	86	86	**71.75**
Qwen 3 Text	62	61	39	70	58	36	74	82	78	67.5
Audio Models										
WavLM-Arousal	72	49	29	38	47	54	40	93	86	68.25
Kimi Audio	69	92	63	64	**72**	54	73	58	98	70.75
Gemini	68	83	48	84	70.75	40	91	66	69	66.5
Qwen 2 Audio	65	71	35	73	61	46	95	83	84	77
Aero	67	69	31	72	59.75	49	87	91	87	**78.50**
Precision and Recall for Non-Negative Emotion										
	Precision				**Avg. Precision**	**Recall**				**Avg. Recall**
Models	**ProdEval**	**CueEval**	**MELD**	**IEMOCAP**		**ProdEval**	**CueEval**	**MELD**	**IEMOCAP**	
Text Models										
BERT	71	74	78	49	68	88	66	82	77	78.25
Llama 3.1	72	61	92	66	72.75	86	82	54	50	**68**
Qwen 2 Text	72	73	92	67	**76**	85	54	50	43	58
Qwen 3 Text	71	75	91	62	74.75	88	62	57	52	64.75
Audio Models										
WavLM-Arousal	78	58	91	96	80.75	89	66	24	69	62
Kimi Audio	77	81	87	98	**85.75**	87	95	89	61	**83**
Gemini	73	92	87	64	79	90	85	76	81	**83**
Qwen 2 Audio	75	94	90	71	82.5	86	68	49	56	64.75
Aero	76	87	92	73	82.00	86	68	33	52	59.75

Table 4: The prediction performance of the audio- and text-based baseline models and our finetuned audio-based SLM and distilled audio-based SLM on different test sets.

of the text- and audio-based models are shown in Table 4 which we will discuss in detail in the following sections.

5.2.2 Performance of text-based models

We observe that among the text-based models Llama 3.1 has the highest avg. Accuracy and ROC-AUC, whereas in terms of recall of negative emotion, Qwen 2 Text has the highest avg. recall. Qwen 3 Text has avg. accuracy, avg. ROC-AUC and avg. recall of negative emotion, which is closer to Llama 3.1. However, interestingly Qwen 3 Text has only 4B parameters compared to Llama 3.1 which has 8B parameters. Since, Qwen 3 models are trained on a combination of COT generated reasoning data and curated datasets, so the implicit thinking ability is ingrained in the model that may contribute to improved performance even when the model has less parameters. Finally, we observe that BERT is the worst model in terms of recall of negative emotion. However, in terms of avg. precision, the performance of BERT is comparable to that of other text-based models. In the case of non-negative emotion, the avg. precision and recall of all the text-based models is comparable, which is expected since non-negative emotion is the majority class and is easier to detect compared to negative emotion labels.

5.2.3 Performance of audio-based models

We first discuss the results of our finetuned Qwen 2 Audio model and other audio models that we use as baselines. We observe that our finetuned Qwen 2 Audio model, surpasses the text-based models by considerable margins in terms of avg. accuracy, avg. ROC-AUC and avg. recall on negative emotion. WavLM Arousal model gives the best performance in ProdEval but does not perform well on CueEval as the dataset is based on linguistic cues, while the model is more inclined towards detecting acoustic cues. Kimi Audio performs better

than Qwen 2 Audio even at zero-shot because Kimi Audio improves upon the Qwen 2.5 7B architecture by using Whisper and GLM tokenizer to tokenize the audio which may play a role in improving the performance. We also observe that Gemini gives performance that is comparable to Qwen 2 Audio in ProdEval, while it gives a higher performance in CueEval. Thus, showing that when using acoustic and linguistic cues, Gemini will require more context to give better results, but it will perform well when the dataset has only linguistic cues.

In MELD, the performance of Qwen 2 Audio degrades in terms of accuracy, ROC-AUC and precision. This could be because MELD is based on TV series and the emotions tend to be exaggerated and dramatic which may have negatively affected the model. In case of IEMOCAP, the accuracy and ROC-AUC of Qwen 2 Audio is comparable to Gemini, while the recall on negative emotion outperforms Gemini. Since, IEMOCAP comprises of conversation between multiple speakers and consists of genuine emotions, it is closer to our ProdEval and CueEval datasets. Thus, the results on ProdEval, CueEval and IEMOCAP datasets show that SLMs that are fine-tuned for domain-specific tasks can outperform generalized larger LLMs. Similar, to text-based models, in the case of non-negative emotion, the avg. precision and recall of all the audio-based models is comparable. Overall, we can conclude that the audio-based models outperform the text-based models as they factor in the linguistic and acoustic cues present in the utterance to predict the emotion.

5.2.4 Performance of distilled audio-based model

The Aero model is distilled from our finetuned Qwen 2 Audio model. The results show that the distilled Aero model has the highest recall on negative emotion among all the text- and audio based models. Since Aero is distilled from our finetuned Qwen 2 Audio, it follows the similar trends as Qwen 2 Audio across all the datasets. The avg. accuracy of Aero is slightly lower than that of Qwen 2 Audio. At dataset level, the recall of negative emotion, is consistently better than Qwen 2 Audio and also exceeds or remains on par with Kimi Audio and Gemini. Thus, we show that distillation from teacher model to student model can improve classification performance even if both models belong to different language model families. Overall, in production, we need a high-recall model for detecting negative emotion and Aero fulfills that requirement.

Models	CPU Latency (in seconds)
BERT	0.04
Qwen 2 Text	5.56
WavLM Arousal	0.66
Qwen 2 Audio	7.36
Aero	**2.56**

Table 5: Inference time CPU latency for text-based models and audio-based models

5.3 Latency Analysis

As we need our model to have a high-recall as well as low latency during inference time, so we analyze the inference time CPU latency of Qwen 2 Audio and Aero. We compare these models with Qwen 2 Text, which is the best performing text-based model in terms of recall of negative emotion, BERT and WavLM-Arousal, which are light-weight models and have only a few million parameters. The average latency of the models is shown in Table 5. We observe that avg. latency of BERT and WavLM-Arousal is less than 1 second because both models have low parameters. However, the classification performance is much lower than the other SLMs discussed earlier. Qwen 2 Text and Qwen 2 Audio have avg. latency of 5.56s and 7.36s, respectively, which make them unsuitable for real-time deployment. However, Aero has an avg. latency of 2.56 secs while maintaining a classification performance that is almost on par with Qwen 2 Audio and outperforms other models in terms of recall of negative emotion. Thus, it is suitable for real-time deployment while maintaining quality performance.

6 Conclusions

We developed low-latency, high-recall audio-based SLM for real-time detection of negative emotions in customer care, with consistent performance across internal and public test sets. Using LLMs for large-scale annotation, we created realistic training data with quality close to human labels. Audio-based models outperformed text-only approaches, effectively capturing emotion from short utterances without context. Notably, smaller fine-tuned models matched or exceeded the performance of larger LLMs. For future work, we aim to further improve latency and recall, and to monitor system performance post-deployment in real-world settings.

Limitations

Despite the comprehensive evaluation, several limitations should be noted. First, while we focus on internal datasets that closely represent real-world data, their scope may still be limited in diversity, which could affect the generalizability of our results to all production environments. Additionally, our approach primarily targets negative emotion detection, and its performance in recognizing other emotional states remains an area for further exploration. Finally, while we aim for real-time applicability, factors such as computational resources and model scalability in dynamic environments were not fully addressed in this study.

References

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

Nicolas Boizard, Kevin El Haddad, CELINE HUDELOT, and Pierre Colombo. Towards cross-tokenizer distillation: the universal logit distillation loss for llms. *Transactions on Machine Learning Research*.

Joost Broekens, Bernhard Hilpert, Suzan Verberne, Kim Baraka, Patrick Gebhard, and Aske Plaat. 2023. Fine-grained affective processing capabilities emerging from large language models. In *Proceedings of the 11th International Conference on Affective Computing and Intelligent Interaction (ACII 2023)*, pages 1–8, Cambridge, MA, USA. IEEE.

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

Mireia Hernandez Caralt, Ivan Sekulić, Filip Carevic, Nghia Khau, Diana Nicoleta Popa, Bruna Guedes, Victor Guimarães, Zeyu Yang, Andre Manso, Meghana Reddy, and 1 others. 2025. "stupid robot, i want to speak to a human!" user frustration detection in task-oriented dialog systems. In *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, pages 276–285.

Youjun Chen, Xurong Xie, Haoning Xu, Mengzhe Geng, Guinan Li, Chengxi Deng, Huimeng Wang, Shujie Hu, and Xunying Liu. 2025. Towards llm-empowered fine-grained speech descriptors for explainable emotion recognition. *arXiv preprint arXiv:2505.23236*.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. 2024. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*.

Tim Dettmers, Mike Lewis, Sam Shleifer, and Luke Zettlemoyer. 2021. 8-bit optimizers via block-wise quantization. *arXiv preprint arXiv:2110.02861*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Soumya Dutta and Sriram Ganapathy. 2025. Llm supervised pre-training for multimodal emotion recognition in conversations. *arXiv preprint arXiv:2501.11468*.

Yumeng Fu. 2024. Ckerc: Joint large language models with commonsense knowledge for emotion recognition in conversation. *arXiv preprint arXiv:2403.07260*.

Apoorva Garg, Sagnik Ghosh, Poonam Ramu, Naren Vempaty, and Harsh Shah. 2024. Deep learning approaches for speech emotion recognition: Challenges and opportunities. *ICT Express*.

L. Goncalves, A. N. Salman, A. Reddy Naini, L. Moro-Velazquez, T. Thebaud, L. Paola Garcia, N. Dehak, B. Sisman, and C. Busso. 2024. Odyssey2024 - speech emotion recognition challenge: Dataset, baseline framework, and results. In *Odyssey 2024: The Speaker and Language Recognition Workshop)*, volume To appear, Quebec, Canada.

Taesik Gong, Josh Belanich, Krishna Somandepalli, Arsha Nagrani, Brian Eoff, and Brendan Jou. 2023. Lanser: Language-model supported speech emotion recognition. In *Proceedings of Interspeech 2023*.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, Xin Zhou, Enzhi Wang, and Xiaohang Dong. 2024. Better zero-shot reasoning with role-play prompting. In *NAACL-HLT*.

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

Hsien-Chin Lin, Shutong Feng, Christian Geishauser, Nurul Lubis, Carel van Niekerk, Michael Heck, Benjamin Ruppik, Renato Vukovic, and Milica Gasić. 2023. Emous: Simulating user emotions in task-oriented dialogues. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2526–2531.

Yi-Cheng Lin, Haibin Wu, Huang-Cheng Chou, Chi-Chun Lee, and Hung-yi Lee. 2024. Emo-bias: A large scale evaluation of social bias on speech emotion recognition. *arXiv preprint arXiv:2406.05065*.

Fei Ma, Yucheng Yuan, Yifan Xie, Hongwei Ren, Ivan Liu, Ying He, Fuji Ren, Fei Richard Yu, and Shiguang Ni. 2025. Generative technology for human emotion recognition: A scoping review. *Information Fusion*, 115:102753.

Hui Ma, Jian Wang, Hongfei Lin, Bo Zhang, Yijia Zhang, and Bo Xu. 2023. A transformer-based model with self-distillation for multimodal emotion recognition in conversations. *IEEE Transactions on Multimedia*.

Liyizhe Peng, Zixing Zhang, Tao Pang, Jing Han, Huan Zhao, Hao Chen, and Björn W. Schuller. 2024. Customising general large language models for specialised emotion recognition tasks. In *Proceedings of the 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11326–11330, Seoul, South Korea. IEEE.

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

Yuntao Shou, Wei Ai, Jiayi Du, Tao Meng, Haiyan Liu, and Nan Yin. 2024. Efficient long-distance latent relation-aware graph neural network for multi-modal emotion recognition in conversations. *arXiv preprint arXiv:2407.00119*.

Armand Stricker and Patrick Paroubek. 2024. A unified approach to emotion detection and task-oriented dialogue modeling. *arXiv preprint arXiv:2401.13789*.

Kimi Team. 2024. Kimi-audio technical report. *Preprint*, arXiv:arXiv:placeholder.

Qwen Team and 1 others. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2:3.

Shreya G Upadhyay, Woan-Shiuan Chien, and Chi-Chun Lee. 2025. Is it still fair? investigating gender fairness in cross-corpus speech emotion recognition. *arXiv preprint arXiv:2501.00995*.

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

Yuezhou Wu, Siling Zhang, and Pengfei Li. 2025. Multi-modal emotion recognition in conversation based on prompt learning with text-audio fusion features. *Scientific Reports*, 15(1):5584.

Yaoxun Xu, Hangting Chen, Jianwei Yu, Qiaochu Huang, Zhiyong Wu, Shi-Xiong Zhang, Guangzhi Li, Yi Luo, and Rongzhi Gu. 2024. Secap: Speech emotion captioning with large language model. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence (AAAI-24)*, pages 19323–19331.

Jieying Xue, Minh-Phuong Nguyen, Blake Matheny, and Le-Minh Nguyen. 2024. Bioserc: Integrating biography speakers supported by llms for erc tasks. In *International Conference on Artificial Neural Networks*, pages 277–292. Springer.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Yazhou Zhang, Mengyao Wang, Youxi Wu, Prayag Tiwari, Qiuchi Li, Benyou Wang, and Jing Qin. 2023. Dialoguellm: Context and emotion knowledge-tuned large language models for emotion recognition in conversations. *arXiv preprint arXiv:2310.11374*.

Zixing Zhang, Liyizhe Peng, Tao Pang, Jing Han, Huan Zhao, and Björn W. Schuller. 2024. Refashioning emotion recognition modelling: The advent of generalised large models. *IEEE Transactions on Computational Social Systems*, 11(5):6690–6704.

A Prompts used for using Gemini 2.0 Flash for annotation

You are given the conversation between a virtual agent and user. The question asked by the virtual agent is in text format and the utterance of the user is in audio format.

Definitions:
Negative Emotion is expressed through anger, irritation, dissatisfaction, etc. The cues that show negative emotion could be linguistic, acoustic, or both. Linguistic cues include negative/angry words, swearing, complaining about the system or its suggestions, repeating the same request multiple times, etc. Acoustic cues include changes in the tone of the speaker or other voice changes including: a higher pitch, louder volume, more tension or strain which can sound tight or harsh, or sudden changes in loudness or speed.

Non-Negative Emotion: Mark an utterance as non-negative if the caller does not seem to have a negative emotion in that utterance. Since we are doing a binary labeling here, positive emotions will also be marked as non-negative.
Note that sometimes negative emotion can be more subtle and does not always involve negative language.

We are looking for the point in the conversation where the caller is starting to be frustrated with the system. This can happen for different reasons, including but not limited to:

1. Customer makes multiple requests wanting to talk to a live agent. If the caller is asking to talk to a live agent then check the context to see if the caller has asked multiple times, if not check whether there are some acoustic cues which reflect negative emotion while asking for a live agent.

2. System misunderstands the caller and repeats the same options or reprompts multiple times

3. Caller does not understand or does not have the information that virtual assistant is asking

4. Long, unresolved conversations where the user's task remains incomplete.

Return only "Negative Emotion" or "Non-Negative Emotion" based on the audio. Do not add any additional comments.

Figure 2: Prompt used for Simple prompting approach

You are given the conversation between a virtual agent and user. The question asked by the virtual agent is in text format and the utterance of the user is in audio format.

Definitions:
Negative Emotion is expressed through anger, irritation, dissatisfaction, etc. The cues that show negative emotion could be linguistic, acoustic, or both. Linguistic cues include negative/angry words, swearing, complaining about the system or its suggestions, repeating the same request multiple times, etc. Acoustic cues include changes in the tone of the speaker or other voice changes including: a higher pitch, louder volume, more tension or strain which can sound tight or harsh, or sudden changes in loudness or speed.

Non-Negative Emotion: Mark an utterance as non-negative if the caller does not seem to have a negative emotion in that utterance. Since we are doing a binary labeling here, positive emotions will also be marked as non-negative.
Note that sometimes negative emotion can be more subtle and does not always involve negative language.

We are looking for the point in the conversation where the caller is starting to be frustrated with the system. This can happen for different reasons, including but not limited to:

1. Customer makes multiple requests wanting to talk to a live agent. If the caller is asking to talk to a live agent then check the context to see if the caller has asked multiple times, if not check whether there are some acoustic cues which reflect negative emotion while asking for a live agent.

2. System misunderstands the caller and repeats the same options or reprompts multiple times

3. Caller does not understand or does not have the information that virtual assistant is asking

4. Long, unresolved conversations where the user's task remains incomplete.

Let's think step by step to understand the emotion of the utterance of the user. Provide some explanation.

Figure 3: Prompt used for ZCOT_Simple, ZCOT_Context, and ZCOT_Detailed_Context prompting approaches

You are given the conversation between a virtual agent and user. The question asked by the virtual agent is in text format and the utterance of the user is in audio format.

Definitions:
Negative Emotion is expressed through anger, irritation, dissatisfaction, etc. The cues that show negative emotion could be linguistic, acoustic, or both. Linguistic cues include negative/angry words, swearing, complaining about the system or its suggestions, repeating the same request multiple times, etc. Acoustic cues include changes in the tone of the speaker or other voice changes including: a higher pitch, louder volume, more tension or strain which can sound tight or harsh, or sudden changes in loudness or speed.

Non-Negative Emotion: Mark an utterance as non-negative if the caller does not seem to have a negative emotion in that utterance. Since we are doing a binary labeling here, positive emotions will also be marked as non-negative.
Note that sometimes negative emotion can be more subtle and does not always involve negative language.

We are looking for the point in the conversation where the caller is starting to be frustrated with the system. This can happen for different reasons, including but not limited to:

1. Customer makes multiple requests wanting to talk to a live agent. If the caller is asking to talk to a live agent then check the context to see if the caller has asked multiple times, if not check whether there are some acoustic cues which reflect negative emotion while asking for a live agent.

2. System misunderstands the caller and repeats the same options or reprompts multiple times

3. Caller does not understand or does not have the information that virtual assistant is asking

4. Long, unresolved conversations where the user's task remains incomplete.

The sequence number of the utterances in the conversation is given in brackets as Utterance sequence no.
The dialog state id indicates the dialog state of the user utterance is provided for each user utterance in brackets as Dialog state id.
Let's first understand the problem and devise a plan to solve the problem. Then, let's carry out the plan and solve the problem step by step.
Pay attention to the context of the conversation (if available), use the text of the context, utterance sequence numbers and dialog state ids to understand the current utterance. Use this to refine your plan. Pay attention to commonsense and logical reasoning.

Figure 4: Prompt used for PAS_Detailed_Context prompting approach

B Impact of Context Window Size on Annotation Performance

Context Window Size	ROC-AUC	Accuracy
1	64.31	69.38
2	61.77	70.51
3	**64.48**	**70.82**
4	64.32	70.01
5	63.64	69.69
Negative Emotion		
Context Window Size	**Precision**	**Recall**
1	59	47
2	68	32
3	63	**43**
4	61	45
5	60	43
Non-Negative Emotion		
Context Window Size	**Precision**	**Recall**
1	74	82
2	71	92
3	73	86
4	73	84
5	73	85

Table 6: Classification performance of Simple prompting approach for different context window sizes. The context window size indicates the number of previous agent-customer interactions that are provided as context.

C Prompt used in Audio-based SLMs

Prompt for audio-based SLMs
`You are a helpful assistant.`
`User: Detect the emotion from the audio. Classify as Negative or Non-Negative emotion`
`Audio: example_audio.wav`
`Assistant: <Emotion>`

Table 7: Prompt template used for Aero, Qwen 2 Audio, Kimi Audio, Gemini models

D Prompt used in Text-based SLMs

Prompt for text-based SLMs
`You are a helpful assistant.`
`User: Detect the emotion from the text. Classify as Negative or Non-Negative emotion`
`Text: <Customer utterance text>`
`Assistant: <Emotion>`

Table 8: Prompt template used for Llama 3.1, Qwen 2 Text, Qwen 3 Text models