Beyond Classification: Towards Speech Emotion Reasoning with Multitask AudioLLMs

Wenyu Zhang[1], Yingxu He[1*], Geyu Lin[1*], Zhuohan Liu[1*], Shuo Sun[1*], Bin Wang[1*], Xunlong Zou[1*], Jeremy H. M. Wong[1], Qiongqiong Wang[1], Hardik B. Sailor[1], Nancy F. Chen[1,2], Ai Ti Aw[1]

[1]Institute for Infocomm Research (I^{2}R), Agency for Science, Technology and Research
[2]Centre for Frontier AI Research (CFAR), Agency for Science, Technology and Research
Singapore

Abstract

Audio Large Language Models (AudioLLMs) have achieved strong results in semantic tasks like speech recognition and translation, but remain limited in modeling paralinguistic cues such as emotion. Existing approaches often treat emotion understanding as a classification problem, offering little insight into the underlying rationale behind predictions. In this work, we explore emotion reasoning, a strategy that leverages the generative capabilities of AudioLLMs to enhance emotion recognition by producing semantically aligned, evidence-grounded explanations. To support this in multitask AudioLLMs, we introduce a unified framework combining reasoning-augmented data supervision, dual-encoder architecture, and task-alternating training. This approach enables AudioLLMs to effectively learn different tasks while incorporating emotional reasoning. Experiments on IEMOCAP and MELD show that our approach not only improves emotion prediction accuracy but also enhances the coherence and evidential grounding of the generated responses. Experiments on two out-of-domain datasets demonstrate the generalization capabilities of the resulting model.

1 Introduction

Recent advancements in Audio Large Language Models (AudioLLMs) (MERaLiON Team, 2024; Tang et al., 2024; Chu et al., 2023; Hu et al., 2023; Das et al., 2024; D'efossez et al., 2024) have driven significant progress in spoken language understanding, particularly for tasks focused on semantic content such as automatic speech recognition (ASR), speech translation (ST), and spoken question answering (SQA). These models typically rely on large-scale audio-text alignment to align spoken inputs with textual outputs (Ji et al., 2024; Held et al., 2024). However, current AudioLLMs are limited in modeling paralinguistic information, such as emotion, which is crucial for applications requiring emotionally aware or empathetic machine behavior (Wang et al., 2024a; Sakshi et al., 2025; Ao et al., 2024).

Traditional emotion recognition approaches in speech primarily focus on categorical classification (e.g., predicting whether a speaker is angry or sad) (Ma et al., 2024; Fu et al., 2025; Zhao et al., 2025). While effective for high-level emotion detection, such methods offer little interpretability or reasoning about why an emotion is being expressed.

In this work, we leverage the generative capabilities of AudioLLMs to incorporate reasoning (Ma et al., 2025; Xie et al., 2025) as a means to improve emotion recognition. Rather than treating emotion understanding as a purely discriminative task, we guide models to generate grounded, semantically aligned explanations that reflect both what is said (semantic content) and how it is said (paralinguistic cues). We categorize emotion recognition outputs into three types in Figure 1: (1) **Label Only**: direct classification (e.g. "The speaker is feeling angry"), with no explanation or grounding; (2) **Interpretive Reasoning**: explanation via paraphrased intent or inferred state (e.g. expressing frustration due to repeated failure); (3) **Evidence-Grounded Reasoning**: the most desirable form, which combines emotion labels with quoted utterances (e.g. "I'm not starting over again") and interprets them to justify the emotional state.

To this end, we propose a new multitask AudioLLM framework with multi-faceted strategies across data, architecture, and training. To guide the model's generative capabilities, we construct reasoning-augmented supervision signals from transcript-aligned data, allowing the model to learn to produce emotion explanations grounded in both linguistic and paralinguistic evidence. Our model architecture employs a dual-encoder design that disentangles semantic and emotional representations. We also propose a task-alternating training strategy that separately optimizes the semantic and

Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 1132–1148

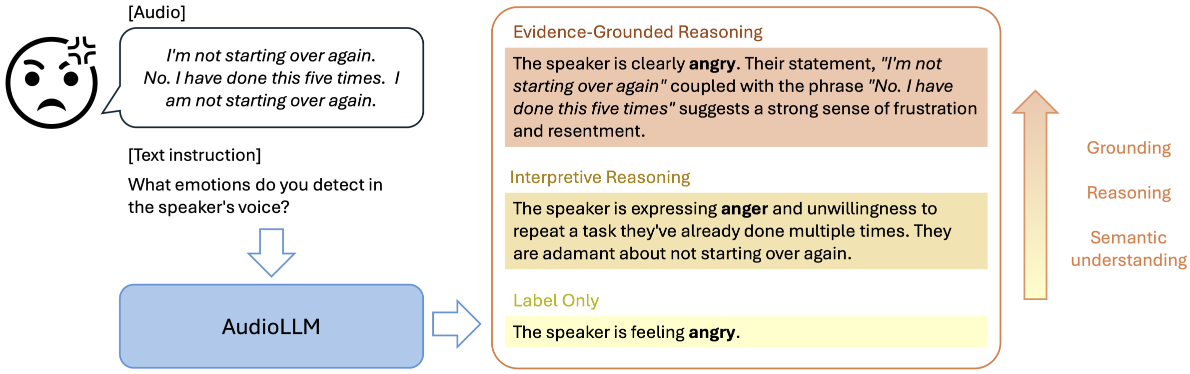

Figure 1: Overview: Our proposed method leverages the generative capabilities of AudioLLM to go beyond classification, producing emotion labels alongside grounded, transcript-informed explanations that reflect the semantic and paralinguistic content of the input speech.

emotion encoders on their respective objectives, aiming to balance performance across tasks. Our framework is evaluated on benchmark datasets for emotion and sentiment recognition, namely IEMOCAP (Busso et al., 2008) and MELD (Poria et al., 2019), as well as ASR and SQA tasks. In summary, our main contributions are:

- We propose a reasoning-augmented approach for speech emotion recognition, enabling AudioLLMs to generate semantically aligned, evidence-grounded explanations that enhance both interpretability and prediction accuracy.
- We introduce a unified framework with multi-faceted strategies in data construction (reasoning target creation), architecture (dual-encoder design), and training (task-alternating training) for multitask AudioLLMs.
- We conduct comprehensive experiments, showing that our approach effectively balances different task performances, improves emotion predictions with minimal effects on other tasks, and enables the coherence and grounding of generated emotion reasoning.

2 Related Works

2.1 AudioLLMs

Multimodal large language models (LLMs), including AudioLLMs (MERaLiON Team, 2024; Tang et al., 2024; Chu et al., 2023; Deshmukh et al., 2023; Hu et al., 2023; Das et al., 2024), commonly adopt a modular architecture comprising three core components: (1) a modality-specific encoder that extracts features from non-textual inputs, (2) a projection or adapter module that maps these features into a representation space compatible with the LLM's tokenizer, and (3) a pretrained LLM that generates free-form text responses based on the projected modality tokens and natural language prompts. For instance, Qwen-Audio (Chu et al., 2023) connects a Whisper-large-v2 (Radford et al., 2023) speech encoder to the Qwen-7B (Bai et al., 2023) language model. To capture richer audio representations, several models employ dual encoders that separately model semantic and acoustic information. SALMONN (Tang et al., 2024) integrates Whisper-large-v2 and BEATs (Chen et al., 2023) with Vicuna-13B (Chiang et al., 2023), while WavLLM (Hu et al., 2023) utilizes Whisper-large-v2 and WavLM-base (Chen et al., 2022), interfaced with LLaMA-2-chat-7B (Touvron et al., 2023).

Distillation approaches use LLMs to generate responses from speech transcriptions or metadata, such as gender and emotion, to train AudioLLMs. Kang et al. (2024) uses an LLM to generate responses to expressive speech prompts, Wang et al. (2024b) generate emotion-aware text continuations, and Lu et al. (2024a,b) generate detailed captions that reflect writing styles and tones.

Recent works explore enabling AudioLLMs to reason. Audio-CoT (Ma et al., 2025) evaluates training-free chain-of-thought prompting, which requires AudioLLMs capable of general instruction following. Audio-Reasoner (Xie et al., 2025) trains with a structured reasoning framework consisting of planning, captioning, reasoning and summarizing steps. (Li et al., 2025a) argues that the complex reasoning process in Audio-Reasoner may not be necessary, and the best practice remains an open research question.

2.2 Emotion recognition in AudioLLMs

Recent research in emotion recognition within AudioLLMs has explored a variety of strategies to

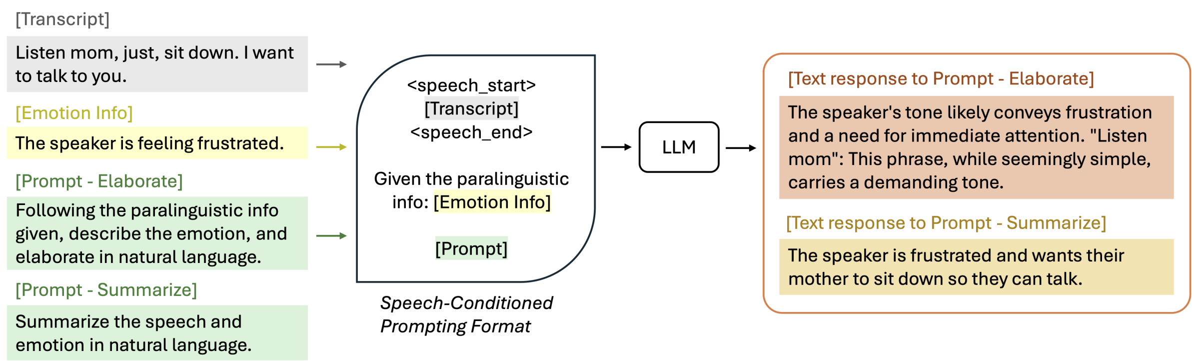

Figure 2: Emotion reason extraction: We input the transcript, emotion label, and reasoning prompt into a speech-conditioned prompting template to elicit grounded and semantically aligned emotion explanations from the LLM. The *Summarize* prompt encourages interpretive reasoning based on the implied context, while the *Elaborate* prompt encourages evidence-grounded reasoning based on explicit cues from the transcript. For more coarse-grained sentiment reason extraction, replace the word "emotion" with "sentiment (positive, negative, neutral)" in the reasoning prompt.

enhance affective understanding from speech (Bellver et al., 2024). These approaches leverage conversational context, paralinguistic cues, and ASR-generated transcripts to improve recognition accuracy. For instance, Sun et al. (2024) employs ASR and LLMs in a cascaded pipeline to transcribe and analyze emotional content, though such pipelines are susceptible to error propagation. SECap (Xu et al., 2024) adopts contrastive and mutual information learning to disentangle semantic and emotional representations in speech. Fu et al. (2025) model speaker traits by prompting LLMs to infer emotional states based on listener responses. C²SER (Zhao et al., 2025) combines Whisper and Emotion2Vec encoders with Chain-of-Thought prompting to inject contextual reasoning into emotion classification. SpeechCueLLM (Wu et al., 2024) introduces descriptive cues, such as volume, pitch and speaking rate, into prompts to enrich LLM inputs with prosodic information. (Li et al., 2024) improves emotion recognition on ASR transcripts by revising transcription errors.

In contrast to prior works, which focus primarily on improving emotion classification accuracy through architectural or input-level enhancements, our approach shifts the paradigm toward emotion reasoning. Rather than outputting a single emotion label, we leverage the generative capabilities of AudioLLMs to produce semantically coherent, evidence-grounded explanations.

2.3 Traditional and Deep Learning Approaches to Emotion Recognition

Research on speech emotion recognition are grounded in the analysis of acoustic–prosodic cues such as pitch, energy, and spectral dynamics as indicators of affective state (Scherer, 2003; Ververidis and Kotropoulos, 2006). The field established standardized low-level descriptors with the Geneva Minimalistic Acoustic Parameter Set (GeMAPS) to promote reproducibility and interpretability in affective computing (Eyben et al., 2015). With the advent of deep learning, convolutional and recurrent architectures supplanted traditional classifiers, learning hierarchical temporal and spectral representations directly from audio (Trigeorgis et al., 2016; Neumann and Vu, 2017; Zhang et al., 2018). Transformer-based and self-supervised speech models have since advanced the state of the art (Pepino et al., 2021; Wagner et al., 2023; Mai et al., 2024). Despite substantial progress, the community continues to emphasize the tension between the growing predictive power of emerging AudioLLMs and the ongoing need to understand how acoustic features and learned representations encode affective information, a challenge that also motivates our work.

3 Proposed Method

We propose a dual-encoder multitask AudioLLM framework that jointly models speech content and emotional reasoning. Our architecture integrates a general-purpose speech encoder and a specialized emotion-centric encoder, which are then connected to a large language model (LLM). To facilitate rich supervision, we introduce reasoning-augmented training targets derived from speech transcripts and emotion labels. Additionally, we adopt a task-alternating training strategy to ensure modular specialization and effective fusion of complementary

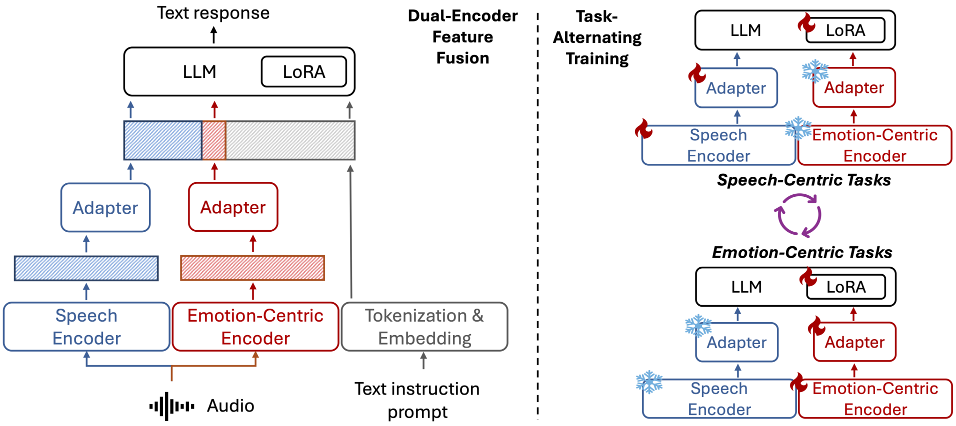

Figure 3: Dual-encoder feature fusion and task-alternating training: We combine features from a general-purpose speech encoder and an emotion-centric encoder. Each encoder and its adapter are trained by alternating between speech-centric and emotion-centric tasks, enabling effective multitask learning with disentagled representations.

features.

3.1 Emotion reason extraction

We introduce reasoning-augmented training targets that pair discrete emotion labels with natural language explanations. These explanations are derived through a prompting-based generation procedure, shown in Figure 2. Specifically, we construct a speech-conditioned prompting format that inputs the transcript, its associated emotion label, and a reasoning prompt into an LLM. We employ two distinct prompting strategies: the *Summarize* prompt encourages interpretive reasoning based on the broader implied context, while the *Elaborate* prompt guides the LLM to produce evidence-grounded justifications based on explicit cues from the transcript. The resulting explanations serve as supervision signals that teach the AudioLLM to associate emotional categories with meaningful linguistic and contextual cues, such that the resulting AudioLLM attains more interpretable and context-sensitive emotion understanding.

Using the generated targets, we construct question-answering training data by sampling questions from a curated pool designed to probe emotional understanding. These questions focus on the speaker's affective state and examples include: "How would you interpret the speaker's emotional state from their speech?", "What emotions do you think the speaker is expressing?", and "How would you describe the tone of the speaker's voice?" We apply a similar approach when querying for coarser-grained sentiment, using broader questions that elicit the speaker's overall positive, negative, or neutral disposition.

3.2 Dual-encoder feature fusion

The multitask AudioLLM framework we utilize consists of: (1) a dual-encoder architecture comprising a general-purpose speech encoder E_{speech} and an emotion-centric encoder E_{emotion}, each designed to capture distinct aspects of the audio input; (2) a pair of lightweight adapter modules that project the encoder outputs into a shared latent space; and (3) a pretrained LLM that consumes the fused representation to generate free-form text outputs. The emotion-centric encoder serves to enhance emotion understanding and reasoning capabilities by introducing inductive biases specific to affective cues. An overview of the dual-encoder framework is shown in Figure 3.

We denote the dataset as $(\mathcal{A}, \mathcal{T}, \mathcal{Y})$, where $\mathcal{A}$ is the set of input audio signals, $\mathcal{T}$ is the set of corresponding text instructions, and $\mathcal{Y}$ is the set of output text responses. Given an audio input $a_i \in \mathcal{A}$ for the i-th training sample, we extract two types of audio embeddings: the speech encoder produces $z_i^{\text{speech}} = E_{\text{speech}}(a_i)$, and the emotion-centric encoder produces $z_i^{\text{emotion}} = E_{\text{emotion}}(a_i)$. We investigated different choices of emotion-centric encoder as described in Section 4 and fixed Whisper (Radford et al., 2023), a widely used model for automatic speech recognition (ASR), as the speech encoder throughout our experiments. The utterances are zero-padded to 30 seconds, and the encoder embeddings have sequence length 1500. These encoder embeddings are passed through adapter modules to be reshaped and projected into a shared latent space. The speech encoder embeddings are transformed to have a sequence length of 100, while the emotion encoder embeddings are transformed to have a shorter sequence length of 10, emphasizing condensed emotion-specific representations as a complementary signal with minimal redundancy. We follow MERaLiON-AudioLLM (MERaLiON Team, 2024) in our implementation of the adapter modules: we concatenate the encoder embeddings across multiple time steps to reduce the sequence length, then pass the them through a multilayer perceptron (MLP) with two hidden layers and SiLU activation. The resulting audio token sequences are obtained as $\text{token}_{a_i}^{\text{speech}}$ and $\text{token}_{a_i}^{\text{emotion}}$. We concatenate these audio token sequences across the sequence dimension:

$$\text{token}_{a_i} = \text{token}_{a_i}^{\text{speech}} \oplus_s \text{token}_{a_i}^{\text{emotion}}.$$

Separately, we tokenize the text instruction $t_i \in \mathcal{T}$ as $\text{token}_{t_i} = \text{tokenizer}(t_i)$. The audio and text tokens are then concatenated across the sequence dimension:

$$\text{token}_i = \text{token}_{a_i} \oplus_s \text{token}_{t_i}.$$

Finally, the concatenated tokens are fed into the LLM to generate the target response:

$$\hat{y}_i = \text{LLM}(\text{token}_i).$$

3.3 Task-alternating training

To ensure that each encoder specializes in its respective task, we employ a task-alternating training strategy, as illustrated in Figure 3. Specifically, the speech encoder and its adapter are trained on speech-centric tasks (e.g., spoken question answering, automatic speech recognition), while the emotion-centric encoder and its adapter are trained on emotion-centric tasks (e.g., emotion recognition with explanation generation). In each training round, only the encoder corresponding to the current task, its associated adapter, and the LLM LoRA parameters are updated, while the other encoder and adapter remain frozen. This alternating scheme enables disentangled yet complementary learning of speech and emotion representations. In the final epoch, we update all adapters and the LLM LoRA parameters across all tasks to enhance multimodal alignment.

4 Experiment Setup

4.1 Model implementation

We use Gemma-2-9B-IT (Gemma Team et al., 2024) as the LLM for emotion reason extraction and in the AudioLLM framework. For each encoder, we utilize the encoder module from Whisper-Large-v3 (Radford et al., 2023), which is a popular choice in existing AudioLLMs (Tang et al., 2024; Chu et al., 2023; Hu et al., 2023; MERaLiON Team, 2024). For the emotion-centric encoder, we also experiment with other options such as smaller-sized versions of Whisper, HuBERT (Hsu et al., 2021) and Emotion2Vec (Ma et al., 2024). For each adapter, we use a light-weight multilayer perceptron (MLP) with two hidden layers and SiLU activation function as in MERaLiON-AudioLLM (MERaLiON Team, 2024). We conduct multitask training with batch size 48 for 5 epochs on 8 H100 GPUs, using an AdamW optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ and a learning rate of 5×10^{-5}. The prompt template for LLM input takes the form:

> *"<audio_start> {audio tokens} <audio_end> {text instruction prompt}"*

4.2 Datasets

We conduct training and evaluation on two widely used benchmarks for emotion recognition (ER) and sentiment recognition (SR): IEMOCAP (Busso et al., 2008) and MELD (Poria et al., 2019). The IEMOCAP dataset comprises dyadic conversations between professional actors, where individual utterances are annotated with one of ten categorical emotion labels, namely anger, happiness, neutral, sadness, disgust, fear, surprise, frustration, excited and others. MELD is a multimodal dataset derived from the TV show Friends, containing audio, video, and text for multi-party conversations. In MELD-ER, each utterance is labeled with one of seven emotion classes, namely neutral, joy, disgust, sadness, surprise, anger and fear. We also include MELD-SR for sentiment recognition, where each utterance is labeled as positive, negative, or neutral, to evaluate the model's ability to capture overall sentiment polarity in spoken contexts.

For semantic tasks, we utilize the Spoken Question Answering (SQA) tasks in MNSC (Wang et al., 2025a), a corpus centered on Singlish, a Creole language rooted in English. We select MNSC because the pre-trained encoders and LLM are unlikely to have been exposed to its linguistic patterns during training, reducing the risk of performance bias from prior exposure. For further analysis in Section 6, we experiment with additional SQA tasks, such as Spoken-SQuAD (Lee et al., 2018) and SLUE-P2-SQA5 (Shon et al., 2023), and automatic speech recognition (ASR) tasks such as

MNSC ASR (Wang et al., 2025a) and LibriSpeech (Panayotov et al., 2015).

Further data details and statistics are provided in the Appendix A.1.

4.3 Evaluation

We perform model evaluations on datasets in Section 4.2 using AudioBench (Wang et al., 2024a) and follow its train-test splits to prevent data contamination. ASR tasks are evaluated with word error rate (WER), and remaining ER, SR and SQA tasks are evaluated using LLM-as-a-Judge framework. Model outputs are assessed by Llama-3-70B-Instruct (AI@Meta, 2024) based on given scoring rubrics, and the scores are then normalized to 0-100 scale where higher scores reflect better performance.

For emotion and sentiment recognition, we followed the style of questions employed in the Audiobench evaluation. The questions are phrased to directly elicit the model's identification or interpretation of emotions expressed in speech. Representative examples include: "What emotions do you detect in the speaker's voice?", "How would you interpret the speaker's emotional state from their speech?", and "Based on the speaker's speech patterns, what do you think they are feeling?" We grade each emotion prediction on a binary scale, where a score of 1 indicates semantic alignment with the ground-truth label. Since AudioLLMs generate open-ended responses, traditional metrics such as exact match may be insufficient. The LLM-as-a-Judge approach allows us to assess the factual correctness and relevance of model outputs in a more flexible manner. We analyze the evaluation approach in Section 7.

We extend the LLM-as-a-Judge framework to evaluate the model's evidence-grounded emotion reasoning quality. We extract direct quotes made in the model predictions and assess them using two key metrics: Groundedness Score and Relevance Score. Groundedness assesses how well the model's extracted quotes align with the ground truth transcript, that is, whether they are directly quoted, faithfully paraphrased, or hallucinated. Relevance measures how effectively the extracted quotes support the annotated emotion label. For each prediction, we provide the Llama-3-70B-Instruct model judge with the ground-truth emotion label, speech transcript and extracted quotes, and instruct it to assign a score from 0 to 2 for each criterion based on a given structured scoring rubric.

Training Targets	IEMOCAP	MELD-ER	MELD-SR	Avg
Label Only (Original)	18.6	47.9	48.1	38.2
Interpretive Reasoning	**60.8**	52.6	60.1	57.8
Evidence-Grounded Reasoning	58.6	**54.1**	**61.6**	**58.1**

Table 1: Emotion and sentiment recognition performance of AudioLLM (with Whisper-Large-v3 encoder) trained on different supervision targets. Training with reasoning-augmented targets yields substantial performance improvements.

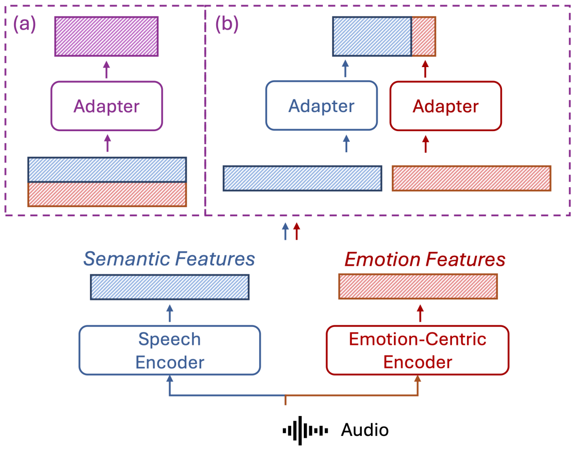

Figure 4: Strategies for feature fusion: We explore (a) fusion along the feature dimension, where features from both encoders are concatenated channel-wise, and (b) fusion along the sequence dimension, where features are concatenated token-wise across time steps.

This evaluation captures both factual alignment and emotional interpretability of the model's output. The evaluation prompt used is provided in the Appendix A.4.

5 Results

5.1 Effectiveness of reasoning-augmented training targets

We evaluate the impact of different supervision targets on emotion and sentiment recognition performance using a baseline AudioLLM equipped with a single Whisper-Large-v3 encoder. From results in Table 1, we observe that models trained with reasoning-augmented targets comprising both emotion labels and natural language explanations consistently outperform those trained with label-only supervision, with almost 20% improvement on average. This indicates that the inclusion of semantically rich and explanatory targets not only enhances interpretability but also leads to substantial gains in recognition capabilities, highlighting the effectiveness of grounding predictions in contextual reasoning. We provide examples of the different model predictions in Appendix A.3.

Concat Dim	Training	ER/SR				SQA					Overall Avg
		IEMOCAP	MELD-ER	MELD-SR	Avg	Part 3	Part 4	Part 5	Part 6	Avg	
None	Joint	43.1	51.9	61.9	52.3	49.4	48.4	57.6	62.4	54.5	53.5
Feature	Joint	40.4	51.5	**62.6**	51.5	47.2	**50.6**	58.6	63.8	55.1	53.5
	Alt 1 epoch	53.7	**54.6**	62.0	**56.8**	49.2	50.0	56.4	63.0	54.7	55.6
	Alt 4 epochs	48.5	52.8	61.4	54.2	42.4	40.4	57.0	62.4	50.6	52.1
Sequence	Joint	44.1	48.8	58.1	50.3	32.8	30.8	40.2	47.6	37.9	43.2
	Alt 1 epoch	**56.6**	52.4	60.7	56.6	48.0	**50.6**	**59.2**	63.4	55.3	55.8
	Alt 4 epochs	55.1	52.0	61.5	56.2	**52.4**	49.4	57.6	**64.4**	**56.0**	**56.1**

Table 2: Performance comparison of different methods for dual-encoder feature fusion and multitask training. Concat Dim "None" indicates the single-encoder baseline. "Joint" indicates that all tasks are trained together on all sets of encoder + adapter for 5 epochs. "Alt x epoch(s)" refers to alternating training of speech-centric and emotion-centric tasks on their respective encoder + adapter every x epoch(s), up to 4 epochs of data, then training all tasks on all adapters at the final epoch.

Training	PL Encoder	ER/SR				SQA					Overall Avg
		IEMOCAP	MELD-ER	MELD-SR	Avg	Part 3	Part 4	Part 5	Part 6	Avg	
Alt 1 epoch	Whisper-Large (637M)	56.6	52.4	60.7	56.6	48.0	50.6	59.2	63.4	55.3	55.8
	Whisper-Small (88M)	56.8	52.5	60.6	56.6	50.8	50.4	58.6	61.4	55.3	55.9
	Whisper-Tiny (8M)	46.2	49.3	53.1	49.5	25.8	23.8	27.0	32.0	27.2	36.7
	HuBERT-XL (962M)	51.3	46.1	51.9	49.8	26.8	23.6	27.0	28.6	26.5	36.5
	Emotion2Vec+ Large (164M)	57.1	50.0	56.6	54.6	28.2	22.4	28.2	32.3	27.8	39.3
	Emotion2Vec+ base (93M)	63.5	45.9	55.0	54.8	24.6	24.2	29.4	29.6	27.0	38.9
Alt 4 epochs	Whisper-Large (637M)	55.1	52.0	61.5	56.2	**52.4**	49.4	57.6	**64.4**	**56.0**	56.1
	Whisper-Small (88M)	55.3	52.6	**61.6**	56.5	48.6	49.2	**61.4**	61.8	55.3	55.8
	Whisper-Tiny (8M)	50.3	52.7	59.7	54.2	47.8	47.0	58.0	62.2	53.8	54.0
	HuBERT-XL (962M)	44.0	49.3	59.0	50.8	49.6	47.8	59.0	62.2	54.7	53.0
	Emotion2Vec+ Large (164M)	**63.8**	**53.0**	61.1	**59.3**	49.2	**51.2**	59.0	62.8	55.6	**57.2**
	Emotion2Vec+ base (93M)	56.3	52.0	60.4	56.2	46.2	45.8	54.4	61.0	51.9	53.7

Table 3: Performance comparison of different choices of emotion-centric encoder in the dual-encoder architecture. "Alt x epoch(s)" refers to alternating training of speech-centric and emotion-centric tasks on their respective encoder + adapter every x epoch(s), up to 4 epochs of data, then training all tasks on all adapters at the final epoch.

5.2 Effectiveness of dual-encoder feature fusion and training

Besides IEMOCAP and MELD, we train and evaluate on MNSC SQA Part 3-6 to assess the effects of our proposed method on non-emotion-centric tasks. Table 2 presents a systematic comparison of key design choices, including model architecture, feature fusion strategies, and task-alternating training. We use Whisper-Large-v3 for both encoders in the dual-encoder architectures. We observe that

1. Dual-encoder architectures can outperform single-encoder baselines, suggesting that incorporating complementary representations enhances overall performance;
2. Concatenation along the sequence dimension yields slightly better results than concatenation along the feature dimension, likely due to better preservation of temporal structure. In the latter, both speech and emotion embeddings are reshaped to length 100, concatenated along the feature dimension, and then passed through a single adapter module. The two types of concatenation are illustrated in Figure 4;
3. Task-alternating training leads to improved performance, particularly for emotion and sentiment recognition, compared to joint multitask training.

In Table 3, we further compare multitask performance across different choices of emotion-centric encoders. We adopt a dual-encoder architecture with features concatenated along the sequence dimension, and apply task-alternating training. Our observations include:

1. Each round of task-specific training must be sufficiently long to ensure model convergence;
2. Encoder selection has a notable impact on performance. Specifically, Emotion2Vec+ Large, which is pre-trained for emotion recognition, provides more relevant features for emotion-centric tasks, leading to improved emotion-centric and overall performance.

We set up training such that the proposed dual-encoder AudioLLM and the vanilla single-encoder AudioLLM are both trained on the amount of data i.e. 5 epochs on the training dataset. The addi-

Scores	IEMOCAP	MELD-ER	MELD-SR	Avg
Quotation	73.8	49.2	49.6	57.5
Groundedness	90.6	79.2	78.5	82.8
Relevance	71.9	60.0	64.1	65.3

Table 4: Evaluation of evidence-grounded reasoning. Quotation Score measures the percentage of predictions containing at least one extractable quote. Groundedness and Relevance scores (0–100 scale) assess alignment with the transcript and support for the ground-truth emotion label, respectively.

tional encoder constitutes only a relatively small portion of the overall architecture. For instance, the encoder module of Whisper-large has 637M parameters, the additional encoder module of Emotion2Vec+ Large has 164M parameters, and the Gemma-2-9B-IT LLM has 9.24B parameters. In practice, the single-encoder AudioLLM achieves a throughput of approximately 1.35 samples/s on a single NVIDIA H100 GPU, whereas the dual-encoder variant runs at approximately 1.26 samples/s.

5.3 Quality of emotion reasoning responses

Table 4 presents the evaluation of evidence-grounded reasoning in the responses generated by our AudioLLM. Across all evaluated datasets, over 49% of the model's responses explicitly include direct quotes from the speech content, serving as supporting evidence for reasoning or interpretation. The model achieves high groundedness scores, averaging 82.8%, indicating that most quotes are faithful to the original speech transcript. Relevance scores average 65.3%, suggesting that the majority of quotes meaningfully support the ground-truth emotion label, though there remains room for improvement, particularly on the MELD dataset.

5.4 Comparison with other models

We compare the emotion and sentiment recognition performance with end-to-end AudioLLMs evaluated in AudioBench: WavLLM (Hu et al., 2023), Qwen2-Audio-7B-Instruct (Chu et al., 2024), Phi-4-Multimodal-Instruct (Abouelenin et al., 2025), MERaLiON-AudioLLM (MERaLiON Team, 2024), Qwen-Audio-Chat (Chu et al., 2023), SALMONN (Tang et al., 2024), R1-AQA (Li et al., 2025b) and Audio-Reasoner (Xie et al., 2025). R1-AQA is trained with reinforcement learning for improved thinking capabilities, and Audio-Reasoner is trained for planning and reasoning. We also compare with cascaded models that process speech in sequential stages by converting audio to text using an automatic speech recognition module before feeding the transcript into a large language model: Whisper-Large-v2 with SEA-LIONv3 (Ng et al., 2025), and Whisper-Large-v3 with Llama-3-8B-Instruct (AI@Meta, 2024). From Table 5, our proposed AudioLLM-Reasoning achieves the best performance for IEMOCAP, MELD-SR and overall.

Model	IEMOCAP	MELD-ER	MELD-SR	Avg
Audio-Reasoner	51.0	**55.9**	54.5	53.8
WavLLM	59.8	41.6	51.1	50.8
Qwen2-Audio	54.0	41.6	53.9	49.8
Cascade: Whisper+SEA-LION	44.3	47.4	56.6	49.4
R1-AQA	57.2	42.8	40.7	46.9
Phi-4-Multimodal	41.0	43.5	51.6	45.4
MERaLiON	48.5	36.4	46.2	43.7
Cascade: Whisper+Llama3	46.7	36.8	45.6	43.0
Qwen-Audio	29.4	50.7	44.9	41.7
SALMONN	23.8	30.5	41.8	32.0
AudioLLM-Reasoning	**63.8**	53.0	**61.1**	**59.3**

Table 5: Emotion and sentiment recognition performance of end-to-end AudioLLMs and cascaded models.

Model	M3ED	CPQA-ER	Avg
Audio-Reasoner	45.2	48.5	46.9
Emotion2Vec+ Large	47.9	37.9	42.9
R1-AQA	38.4	43.1	40.8
AudioLLM-Reasoning	**48.6**	**49.0**	**48.8**

Table 6: Emotion recognition performance on out-of-domain datasets.

5.5 Generalization to out-of-domain datasets

We evaluate the generalizability of the proposed model on two out-of-domain emotion recognition datasets not seen at training: M3ED (Zhao et al., 2022) and CPQA (Wang et al., 2025b). M3ED contains utterances from Mandarin Chinese TV series. CPQA is a contexual paralinguistic question-answering dataset constructed using speech data collected from top Singapore YouTube channels; we use only the emotion recognition set for our evaluation. Both datasets are annotated for seven emotion classes, namely neutral, happy, disgust, sad, surprise, anger and fear. We compare our model with emotion classifier Emotion2Vec+ Large, R1-AQA which is trained with reinforcement learning for improved thinking capabilities, and Audio-Reasoner which is trained for planning and reasoning. From Table 6, our proposed AudioLLM-Reasoning achieves the best performance on both M3ED and CPQA-ER. Additional class-wise comparisons with Emotion2Vec+ Large are in Appendix A.2.

Model	ER/SR (↑)	SQA (↑)	ASR (↓)
Base AudioLLM	44.1	**56.0**	**19.5**
+ Emotion Supervision	**56.4**	54.1	19.6

(a) Base AudioLLM trained on MNSC SQA Part 3-6 and MNSC ASR Part 3-6.

Model	ER/SR (↑)	SQA (↑)	ASR (↓)
Base AudioLLM	35.6	**80.3**	3.8
+ Emotion Supervision	**58.0**	79.0	**3.6**

(b) Base AudioLLM trained on Spoken-SQuAD and SLUE for SQA, and LibriSpeech Clean and Other splits for ASR.

Table 7: Effect of adding emotion supervision on a trained base AudioLLM. Adding emotion supervision improves emotion understanding with slight compromise on the performance of other tasks

6 Further Analysis

We further investigate whether emotion understanding capabilities can be introduced into a model that was not originally trained for these tasks. Starting from a base AudioLLM without any emotion-specific supervision, we explore adding an emotion-centric encoder to the architecture. We train the emotion-centric encoder Emotion2Vec+ Large, adapter and LLM LoRA on emotion-centric tasks, then finetune the adapters and LLM LoRA on all tasks.

Table 7 presents the effect of incorporating emotion supervision into a base AudioLLM trained on different upstream tasks. Across both training configurations, we observe substantial gains in emotion and sentiment recognition, with improvements of +12.3 and +22.4 points, respectively. This enhancement in emotional understanding comes with a slight degradation in the model's performance on the original tasks. For instance, SQA performance remains comparable, and ASR performance is largely preserved or slightly improved. These results highlight that emotional capabilities can be effectively injected into a multimodal model without sacrificing its existing competencies.

7 Assessment of LLM-as-a-Judge metric

We adopt the LLM-as-a-Judge metric for evaluation as the AudioLLMs can generate open-ended and expressive outputs. Traditional metrics relying on exact string matching are insufficient in this context. We identify three main cases where traditional metrics can fail to capture the true quality of model responses: 1. the model uses semantically equivalent but lexically different expressions to describe emotions; 2. the model output includes multiple plausible or related emotional states; and 3. the predicted labels fall into semantically overlapping categories (e.g., excited vs. happy, or anger vs. frustration). Outside of these special cases, the LLM-as-a-Judge metric effectively reduces to an accuracy-like measure, where the model's prediction is compared against a reference by string matching.

Out of the 858 test samples in IEMOCAP, 27.6% of the AudioLLM-Reasoning emotion predictions fall into the special cases. Out of the 2610 test samples in MELD-ER, 6% of the AudioLLM-Reasoning emotion predictions fall into the special cases. To better understand the characteristics of the LLM-as-a-Judge metric, we randomly selected 50 special-case samples each from the IEMOCAP and MELD-ER datasets to be assessed by 4 human evaluators. The evaluators are provided with the audio clips, references and model's answers, and are instructed to rate the model's answers based on their alignment with the audio clips and references. For both IEMOCAP and MELD-ER, only 2% of the selected samples are scored as correct by the LLM judge but scored as wrong by the human evaluators. In contrast, 16% of the IEMOCAP selected samples and 30% of the MELD-ER selected samples are scored as wrong by the LLM judge but scored as correct by the human evaluators, suggesting that the LLM judge tends to be more conservative or strict in its assessments compared to human evaluators.

8 Conclusion

In this work, we propose a unified framework that brings emotion reasoning into multitask AudioLLMs, combining dual encoders, reasoning-augmented supervision, and task-alternating training. Our method improves emotion and sentiment recognition and enables the generation of evidence-grounded explanations, as demonstrated on IEMOCAP and MELD benchmark datasts. This work highlights the potential of generative AudioLLMs for more interpretable and emotionally aware speech understanding.

9 Limitations

While our framework significantly improves both emotion recognition and explanation capabilities in AudioLLMs, several limitations remain. The current benchmarks for emotional reasoning are limited in diversity and scale, highlighting the need for more comprehensive evaluation datasets that capture a wider range of emotional expressions and contextual richness. The explanation generation quality can vary across emotions and speaker styles, especially for subtle or ambiguous affective states. Moreover, the quality of extracted emotion reasoning is dependent on the capabilities of the teacher LLM, which may introduce biases or inaccuracies in supervision. Using weaker or stronger teacher models is expected to affect the quality of the extracted reasoning targets. In this work, however, we set both the teacher LLM and the student LLM initialized within the AudioLLM framework to be Gemma-2-9B-IT. This design choice was made to avoid introducing additional biases from an external teacher and to ensure that our evaluation isolates the contribution of reasoning supervision itself.

10 Ethics Statement

All datasets used in our study, including IEMOCAP and MELD, are publicly available and widely used in the research community. However, we emphasize that caution is necessary when deploying emotion-aware AI systems in real-world or sensitive contexts, as misinterpretation of emotional cues may lead to unintended consequences. Ensuring transparency, user consent, and appropriate safeguards is critical when applying these technologies beyond academic settings.

Acknowledgements

This research is supported by the National Research Foundation, Singapore under its National Large Language Models Funding Initiative. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not reflect the views of the National Research Foundation, Singapore.

References

Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, Dong Chen, Dongdong Chen, Junkun Chen, Weizhu Chen, Yen-Chun Chen, Yiling Chen, Qi Dai, Xiyang Dai, Ruchao Fan, and Xiren Zhou. 2025. Phi-4-Mini technical report: Compact yet powerful multimodal language models via mixture-of-LoRAs. *arXiv*.

AI@Meta. 2024. Llama 3 model card.

Junyi Ao, Yuancheng Wang, Xiaohai Tian, Dekun Chen, Jun Zhang, Lu Lu, Yuxuan Wang, Haizhou Li, and Zhizheng Wu. 2024. SD-Eval: A benchmark dataset for spoken dialogue understanding beyond words. *arXiv*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, and et al. 2023. Qwen technical report. *arXiv*.

Jaime Bellver, Ivan Martín-Fernández, Jose Bravo-Pacheco, Sergio Esteban, Fernando Fernández-Martínez, and Luis D'Haro. 2024. Multimodal audio-language model for speech emotion recognition. In *Odyssey: The Speaker and Language Recognition Workshop*.

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

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, and et al. 2022. WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, Wanxiang Che, Xiangzhan Yu, and Furu Wei. 2023. BEATs: Audio pre-training with acoustic tokenizers. In *ICML*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

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*.

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. 2023. Qwen-Audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv*.

Nilaksh Das, Saket Dingliwal, S. Ronanki, Rohit Paturi, David Huang, Prashant Mathur, Jie Yuan, Dhanush Bekal, Xing Niu, Sai Muralidhar Jayanthi, Xilai Li, Karel Mundnich, Monica Sunkara, Sundararajan Srinivasan, Kyu J Han, and Katrin Kirchhoff. 2024. SpeechVerse: A large-scale generalizable audio language model. *arXiv*.

Alexandre D'efossez, Laurent Mazar'e, Manu Orsini, Am'elie Royer, Patrick P'erez, Herv'e J'egou, Edouard Grave, and Neil Zeghidour. 2024. Moshi: a speech-text foundation model for real-time dialogue. *arXiv*.

Soham Deshmukh, Benjamin Elizalde, Rita Singh, and Huaming Wang. 2023. Pengi: An audio language model for audio tasks. In *NeurIPS*.

Florian Eyben, Klaus R Scherer, Björn W Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y Devillers, Julien Epps, Petri Laukka, Shrikanth S Narayanan, and Khiet P Truong. 2015. The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE Transactions on Affective Computing*, 7(2):190–202.

Yumeng Fu, Junjie Wu, Zhongjie Wang, Meishan Zhang, Lili Shan, Yulin Wu, and Bingquan Liu. 2025. LaERC-S: Improving LLM-based emotion recognition in conversation with speaker characteristics. In *COLING*.

Morgane Riviere Gemma Team, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, and et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv*.

William Held, Minzhi Li, Michael Ryan, Weiyan Shi, Yanzhe Zhang, and Diyi Yang. 2024. Distilling an end-to-end voice assistant without instruction training data. *arXiv*.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 29:3451–3460.

Shujie Hu, Long Zhou, Shujie Liu, Sanyuan Chen, Hongkun Hao, Jing Pan, Xunying Liu, Jinyu Li, Sunit Sivasankaran, Linquan Liu, and Furu Wei. 2023. WavLLM: Towards robust and adaptive speech large language model. *arXiv*.

Shengpeng Ji, Yifu Chen, Minghui Fang, Jialong Zuo, Jingyu Lu, Hanting Wang, Ziyue Jiang, Long Zhou, Shujie Liu, Xize Cheng, Xiaoda Yang, Zehan Wang, Qian Yang, Jian Li, Yidi Jiang, Jingzhen He, Yunfei Chu, Jin Xu, and Zhou Zhao. 2024. WavChat: A survey of spoken dialogue models. *arXiv*.

Wonjune Kang, Junteng Jia, Chunyang Wu, Wei Zhou, Egor Lakomkin, Yashesh Gaur, Leda Sari, Suyoun Kim, Ke Li, Jay Mahadeokar, and Ozlem Kalinli. 2024. Frozen large language models can perceive paralinguistic aspects of speech. *arXiv*.

Chia-Hsuan Lee, Szu-Lin Wu, Chi-Liang Liu, and Hung-yi Lee. 2018. Spoken SQuAD: A study of mitigating the impact of speech recognition errors on listening comprehension. In *Interspeech*.

Gang Li, Jizhong Liu, Heinrich Dinkel, Yadong Niu, Junbo Zhang, and Jian Luan. 2025a. Reinforcement learning outperforms supervised fine-tuning: A case study on audio question answering. *arXiv*.

Gang Li, Jizhong Liu, Heinrich Dinkel, Yadong Niu, Junbo Zhang, and Jian Luan. 2025b. Reinforcement learning outperforms supervised fine-tuning: A case study on audio question answering. *arXiv*.

Yuanchao Li, Yuan Gong, Chao-Han Huck Yang, Peter Bell, and Catherine Lai. 2024. Revise, reason, and recognize: Llm-based emotion recognition via emotion-specific prompts and asr error correction. *ICASSP*.

Ke-Han Lu, Zhehuai Chen, Szu-Wei Fu, He Huang, Boris Ginsburg, Yu-Chiang Frank Wang, and Hung yi Lee. 2024a. DeSTA: Enhancing speech language models through descriptive speech-text alignment. In *Interspeech*.

Ke-Han Lu, Zhehuai Chen, Szu-Wei Fu, Chao-Han Huck Yang, Jagadeesh Balam, Boris Ginsburg, Yu-Chiang Frank Wang, and Hung-yi Lee. 2024b. DeSTA2: Developing instruction-following speech language model without speech instruction-tuning data. In *ICASSP*.

Ziyang Ma, Zhuo Chen, Yuping Wang, Eng Siong Chng, and Xie Chen. 2025. Audio-CoT: Exploring chain-of-thought reasoning in large audio language model. *arXiv*.

Ziyang Ma, Zhisheng Zheng, Jiaxin Ye, Jinchao Li, Zhifu Gao, ShiLiang Zhang, and Xie Chen. 2024. emotion2vec: Self-supervised pre-training for speech emotion representation. In *ACL*.

Trang Mai, Duc Huy Vu, Hoang Nguyen, and Dung Tien Ho. 2024. Dropformer: A dynamic noise-dropping transformer for speech emotion recognition. In *INTERSPEECH*.

MERaLiON Team. 2024. MERaLiON-AudioLLM: Bridging audio and language with large language models. *arXiv*.

Michael Neumann and Ngoc Thang Vu. 2017. Attentive convolutional neural network based speech emotion recognition: Insights on features and data augmentation. In *INTERSPEECH*.

Raymond Ng, Thanh Ngan Nguyen, Yuli Huang, Ngee Chia Tai, Wai Yi Leong, Wei Qi Leong, Xianbin Yong, Jian Gang Ngui, Yosephine Susanto, Nicholas Cheng, Hamsawardhini Rengarajan, Peerat Limkonchotiwat, Adithya Venkatadri Hulagadri, Kok Wai Teng, Yeo Yeow Tong, Bryan Siow, Wei Yi Teo, Wayne Lau, Choon Meng Tan, and 12 others. 2025. SEA-LION: Southeast Asian Languages in One Network. *arXiv*.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An ASR corpus based on public domain audio books. In *ICASSP*.

Luciana Pepino, Paula Riera, and Guillermo Ferroni. 2021. Emotion recognition from speech using wav2vec 2.0 embeddings. In *INTERSPEECH*.

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 *ACL*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *ICML*.

S Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. 2025. MMAU: A massive multi-task audio understanding and reasoning benchmark. In *ICLR*.

Klaus R Scherer. 2003. Vocal communication of emotion: A review of research paradigms. *Speech Communication*, 40(1–2):227–256.

Suwon Shon, Siddhant Arora, Chyi-Jiunn Lin, Ankita Pasad, Felix Wu, Roshan Sharma, Wei Yu Wu, Hung yi Lee, Karen Livescu, and Shinji Watanabe. 2023. SLUE Phase-2: A benchmark suite of diverse spoken language understanding tasks. *ACL*.

Chuanneng Sun, Zeeshan Ahmed, Yingyi Ma, Zhe Liu, Lucas Kabela, Yutong Pang, and Ozlem Kalinli. 2024. Contextual biasing of named-entities with large language models. In *ICASSP*.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. 2024. SALMONN: Towards generic hearing abilities for large language models. In *ICLR*.

Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, and et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv*.

Georgios Trigeorgis, Fabien Ringeval, Rainer Brueckner, Erik Marchi, Mihalis A Nicolaou, Björn W Schuller, and Stefanos Zafeiriou. 2016. Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In *ICASSP*.

Dimitrios Ververidis and Constantine Kotropoulos. 2006. Emotional speech recognition: Resources, features, and methods. *Speech Communication*, 48(9):1162–1181.

Johannes Wagner, Andreas Triantafyllopoulos, Hagen Wierstorf, Maximilian Schmitt, Felix Burkhardt, Florian Eyben, and Björn W Schuller. 2023. Dawn of the transformer era in speech emotion recognition: Closing the valence gap. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10745–10759.

Bin Wang, Xunlong Zou, Geyu Lin, Shuo Sun, Zhuohan Liu, Wenyu Zhang, Zhengyuan Liu, AiTi Aw, and Nancy F Chen. 2024a. AudioBench: A universal benchmark for audio large language models. *NAACL*.

Bin Wang, Xunlong Zou, Shuo Sun, Wenyu Zhang, Yingxu He, Zhuohan Liu, Chengwei Wei, Nancy F Chen, and AiTi Aw. 2025a. Advancing Singlish understanding: Bridging the gap with datasets and multimodal models. *arXiv*.

Chen Wang, Minpeng Liao, Zhongqiang Huang, Junhong Wu, Chengqing Zong, and Jiajun Zhang. 2024b. BLSP-Emo: Towards empathetic large speech-language models. In *EMNLP*.

Qiongqiong Wang, Hardik Sailor, Tianchi Liu, and Ai Ti Aw. 2025b. Contextual paralinguistic data creation for multi-modal speech-llm: Data condensation and spoken qa generation. In *Interspeech*.

Zehui Wu, Ziwei Gong, Lin Ai, Pengyuan Shi, Kaan Donbekci, and Julia Hirschberg. 2024. Beyond silent letters: Amplifying LLMs in emotion recognition with vocal nuances. *arXiv*.

Zhifei Xie, Mingbao Lin, Zihang Liu, Pengcheng Wu, Shuicheng Yan, and Chunyan Miao. 2025. Audio-Reasoner: Improving reasoning capability in large audio language models. *arXiv*.

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 *AAAI*.

Shiqing Zhang, Shiliang Zhang, Tiejun Huang, and Wen Gao. 2018. Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching. *IEEE Transactions on Multimedia*, 20(6):1576–1590.

Jinming Zhao, Tenggan Zhang, Jingwen Hu, Yuchen Liu, Qin Jin, Xinchao Wang, and Haizhou Li. 2022. M3ED: Multi-modal multi-scene multi-label emotional dialogue database. In *ACL*.

Zhixian Zhao, Xinfa Zhu, Xinsheng Wang, Shuiyuan Wang, Xuelong Geng, Wenjie Tian, and Lei Xie. 2025. Steering language model to stable speech emotion recognition via contextual perception and chain of thought. *arXiv*.

A Experiment Details

A.1 Datasets

We conduct training and evaluation on two widely used benchmarks for emotion recognition (ER) and sentiment recognition (SR): IEMOCAP (Busso et al., 2008) and MELD (Poria et al., 2019). IEMOCAP is made available under a custom non-commercial research license, and MELD is distributed under the GNU General Public License v3.0 (GPL-3.0). Since IEMOCAP lacks a predefined train-test split, we adopt the 90-10 split defined in AudioBench (Wang et al., 2024a), with

9035 training samples (anger: 1140, disgust: 2, excited: 1816, fear: 98, frustration: 2608, happiness: 588, neutral: 1539, other: 23, sad: 1120, surprise: 101) and 1004 test samples (anger: 129, disgust: 0, excited: 160, fear: 9, frustration: 309, happiness: 68, neutral: 187, other: 3, sad: 130, surprise: 9). MELD has 9988 training samples (anger: 1109, disgust: 271, fear: 268, joy: 1743, neutral: 4709, sad: 683, surprise: 1205) and 2610 test samples (anger: 345, disgust: 68, fear: 50, joy: 402, neutral: 1256, sadness: 208, surprise: 281).

For semantic tasks, we utilize the Multitask National Speech Corpus (MNSC) (Wang et al., 2025a), specifically SQA Part 3-6 and ASR Part 3-6, released under the Singapore Open Data License. We also use Spoken-SQuAD (Lee et al., 2018) released under CC-BY-SA-4.0 License, SLUE-P2-SQA5 (Shon et al., 2023) which is a collection of datasets released under CC-BY-SA-4.0 License and Apache License 2.0, and LibriSpeech (Panayotov et al., 2015) released under CC-BY-4.0 License.

All experiments in this work respect the respective licenses and usage terms of the datasets.

A.2 Comparison with Emotion2Vec+

We compare the performance of AudioLLM with emotion classifier Emotion2Vec+ Large. For the in-domain dataset MELD-ER in Table 8, AudioLLM trained with reasoning-augmented targets achieves the highest weighted average score (53.0%), outperforming both the label-only variant (47.1%) and Emotion2Vec+ Large (44.7%). As the AudioLLMs are trained on a limited set of emotion datasets (i.e. IEMOCAP and MELD), they tend to fit to the training label distribution. In contrast, Emotion2Vec+ Large is trained on five emotion datasets.

The performance on the out-of-domain dataset CPQA-ER in Table 9 is affected less by the training label distribution. AudioLLM-Reasoning outperforms Emotion2Vec+ Large classifier on 6 out of 7 classes, and significantly outperforms the classifer on both the unweighted average score (45.3% vs. 36.0%) and the weighted average score (49.7% vs. 38.3%). Moreover, the discriminative classifier cannot readily extend beyond its label space. In Table 9, 11 samples involving states such as frustration, embarrassment, and mixture of emotions are excluded. The vanilla AudioLLM trained with label-only targets lacks generalization capabilities and has severely degraded performances.

Class	Num Samples	Emotion2Vec+ Large	AudioLLM - Label Only	AudioLLM - Reasoning
Neutral	1256	54.2	64.7	**83.8**
Joy	402	**54.5**	41.5	36.3
Disgust	68	0.0	8.8	**13.2**
Sadness	208	**32.2**	23.1	15.9
Surprise	281	**38.8**	28.8	26.0
Anger	345	26.1	**32.2**	31.3
Fear	50	2.0	6.0	**8.0**
Unwt Avg	2610	**29.7**	29.3	29.2
Wt Avg	2610	44.7	47.1	**53.0**

Table 8: In-domain comparison: Class-wise emotion recognition performance comparison of Emotion2Vec+ Large classifier vs. AudioLLM (with Emotion2Vec+ Large emotion-centric encoder) trained on Label Only or Reasoning targets on MELD-ER. The unweighted average (Unwt Avg) treats all classes equally regardless of sample size, while the weighted average (Wt Avg) reflects the sample distribution by weighting each class accordingly.

Class	Num Samples	Emotion2Vec+ Large	AudioLLM - Label Only	AudioLLM - Reasoning
Neutral	102	12.7	13.7	**32.4**
Joy	91	71.4	38.5	**74.7**
Disgust	30	13.3	10.0	**20.0**
Sadness	57	**59.6**	22.8	40.4
Surprise	54	31.5	1.9	**33.3**
Anger	83	43.4	28.9	**78.3**
Fear	50	20.0	4.0	**38.0**
Unwt Avg	467	36.0	17.1	**45.3**
Wt Avg	467	38.3	19.7	**49.7**

Table 9: Out-of-domain comparison: Class-wise emotion recognition performance comparison of Emotion2Vec+ Large classifier vs. AudioLLM (with Emotion2Vec+ Large emotion-centric encoder) trained on Label Only or Reasoning targets on CPQA-ER. 11 samples with ground-truth labels outside of the label space of Emotion2Vec+ Large classifier have been excluded.

A.3 AudioLLM responses

We conduct a qualitative analysis of model predictions across different reasoning formats on the IEMOCAP, MELD-ER and MELD-SR datasets. Each example includes the transcript, ground-truth label, and model-generated outputs under three supervision types: label-only, interpretive reasoning, and evidence-grounded reasoning. As shown in Table 10, 11 and 12, label-only responses often fail to capture the correct emotional nuance, defaulting to neutral predictions even when the emotion is apparent. In contrast, interpretive and evidence-grounded reasoning better align with the ground truth, offering richer justifications and improved emotion recognition. Notably, evidence-grounded reasoning demonstrates superior clarity by explicitly linking speech content and affective cues to the predicted emotion.

Example 1	
Transcript: *Is he an actor too, or-?*	
Ground Truth Emotion Label: Neutral	
Label Only: The speaker's speech reflects a neutral state	
Interpretive Reasoning: The speaker is asking if someone was an actor too, implying a shared experience or connection. They seem neutral in their tone, neither excited nor disappointed.	
Evidence-Grounded Reasoning: The speaker's tone is neutral, suggesting they are neither expressing strong positive nor negative emotions. They are simply seeking clarification or information. The phrase "Is he an actor too?" implies a previous context where someone was identified as an actor. The speaker is now inquiring if another person also shares this profession.	
Example 2	
Transcript: *Oh, I will never forget his face. Ooh-rah [LAUGHTER].*	
Ground Truth Emotion Label: Excited	
Label Only: The speaker's speech suggesting a neutral state.	
Interpretive Reasoning: The speaker is laughing and reminiscing about someone's face, implying a positive and humorous memory. They express excitement and amusement.	
Evidence-Grounded Reasoning: The speaker is clearly excited and perhaps a little bit nostalgic. Here's why: "Oh, I will never forget his face": This phrase suggests a strong memory, likely positive given the excited tone. Laughter: Laughter often accompanies happy memories and a sense of lightheartedness. The speaker is reminiscing about someone and a specific moment, and the excitement suggests that the memory is a pleasant one.	
Example 3	
Transcript: *You've got to tell me.*	
Ground Truth Emotion Label: Frustration	
Label Only: The speaker's speech suggesting a neutral state.	
Interpretive Reasoning: The speaker is frustrated and wants the listener to tell them something.	
Evidence-Grounded Reasoning: The speaker is frustrated. They are likely feeling impatient, annoyed, or even angry because they feel like something is not being done or explained properly. The phrase "You've got to tell me" suggests a sense of urgency and a demand for information.	

Table 10: IEMOCAP: Qualitative comparison of emotion predictions. The first three samples in the evaluation dataset are presented. Each example includes the transcript, ground-truth emotion label, and model predictions with their inferred emotions. Emotion labels are in blue, and quotes from the speech content are in orange. Interpretive and evidence-grounded reasoning better align with the ground truth, offering richer justifications and improved emotion recognition.

A.4 Evaluation of emotion reasoning quality

We follow AudioBench's LLM-as-a-Judge framework to evaluate the model's evidence-grounded emotion reasoning quality. We extract direct quotes made in the model predictions and assess them using two key metrics: Groundedness Score and Relevance Score. For each prediction, we provide the Llama-3-70B-Instruct model judge with the ground-truth emotion label, speech transcript and extracted quotes, and instruct it to assign a score from 0 to 2 for each criterion based on a given structured scoring rubric. The scores are then normalized to 0-100 scale. The evaluation prompt used is shown in Figure 5.

Evaluation Prompt Template

```
[Ground Truth Emotion]
{reference}

[Ground Truth Transcript]
{transcript}

[Extracted Quotes from Model Prediction]
{extracted_quotes}

[Evaluation Task]
Evaluate the extracted quotes using the following three criteria.

**Groundedness Score**
Assess whether the extracted quotes are grounded in the ground truth transcript.
Scoring Guide:
Score0: The quotes do not appear in the ground truth transcript and are not
    semantically aligned (i.e., hallucinated or generic).
Score1: The quotes partially match the ground truth transcript. There may be
    loose paraphrasing or selective grounding.
Score2: The quotes are clearly derived from the ground truth transcript, through
     direct quotes or faithful paraphrases.

**Relevance Score**
Assess whether the extracted quotes support the ground truth emotion label.
Scoring Guide:
Score0: The quotes are irrelevant or inconsistent with the ground truth emotion.
     They may even suggest a different emotion.
Score1: The quotes are loosely related to the ground truth emotion but lack
    clarity, specificity, or completeness.
Score2: The quotes clearly and directly support the ground truth emotion.

Respond with the following structured format:

Ground Truth Emotion: (string)
Ground Truth Transcript: (string)
Extracted Quotations from Model Prediction: (list of strings)
Groundedness Score: (int)
Relevance Score: (int)
Explanation: (string - justify the assigned scores)
```

Figure 5: Evaluation prompt used for assessing groundedness and relevance of extracted emotional evidence.

Example 1 **Transcript:** *Yeah, I'm not in that.* **Ground Truth Emotion Label:** Neutral
Label Only: The speaker's emotion seems neutral. **Interpretive Reasoning:** The speaker is denying being in something, and they sound neutral about it. **Evidence-Grounded Reasoning:** The speaker's tone is neutral, indicating they are not experiencing any strong emotions. They are simply stating a fact without any particular emphasis or feeling behind their words.
Example 2 **Transcript:** *He's a dropper!* **Ground Truth Emotion Label:** Anger
Label Only: The speaker's emotion seems neutral. **Interpretive Reasoning:** The speaker is expressing anger and frustration towards someone they perceive as a "dropper," likely someone who is unreliable or lets them down. **Evidence-Grounded Reasoning:** The speaker's tone is neutral, but the statement itself carries a strong implication. While the speaker doesn't express anger or disgust directly, the phrase "He's a dropper!" suggests a negative judgment about the person being referred to. It implies that this person is unreliable, perhaps someone who makes promises they don't keep or fails to deliver on expectations.
Example 3 **Transcript:** *Yes!* **Ground Truth Emotion Label:** Joy
Label Only: The speaker's speech reflects neutral. **Interpretive Reasoning:** The speaker said "Yes." The emotion conveyed is neutral. **Evidence-Grounded Reasoning:** The speaker's tone is neutral, indicating they are not experiencing any strong positive or negative emotions. They are likely calm, composed, and objective in their state of mind.

Table 11: MELD-ER: Qualitative comparison of emotion predictions. The first three samples in the evaluation dataset are presented. Each example includes the transcript, ground-truth emotion label, and model predictions with their inferred emotions. Emotion labels are in blue, and quotes from the speech content are in orange. Interpretive and evidence-grounded reasoning provide richer interpretation of speaker emotion by additionally making use of the speech content.

Example 1 **Transcript:** *Yeah, I'm not in that.* **Ground Truth Emotion Label:** Neutral
Label Only: The speaker's sentiment is neutral. **Interpretive Reasoning:** The speaker is stating that they are not in a particular location or situation. They express this in a neutral tone, indicating no strong feelings about the matter. **Evidence-Grounded Reasoning:** The speaker's tone is neutral, indicating they are not expressing strong positive or negative emotions. They are simply stating a fact without any noticeable inflection or emphasis.
Example 2 **Transcript:** *He's a dropper!* **Ground Truth Emotion Label:** Negative
Label Only: The speaker's sentiment is neutral, based on their speech. **Interpretive Reasoning:** The speaker is expressing a negative sentiment towards someone they perceive as a "dropper," which likely implies a negative trait or action. **Evidence-Grounded Reasoning:** The sentiment is negative. The speaker's tone suggests disapproval or disgust towards the person being referred to as a "dropper." The word "dropper" itself often carries a negative connotation, implying someone who is unreliable, messy, or in some way undesirable.
Example 3 **Transcript:** *Yes!* **Ground Truth Emotion Label:** Positive
Label Only: The speaker's sentiment is neutral, based on their speech. **Interpretive Reasoning:** The speaker said "Yes." The sentiment is neutral. **Evidence-Grounded Reasoning:** The sentiment is positive. The speaker's enthusiastic "Yes!" conveys a strong sense of agreement, excitement, or approval.

Table 12: MELD-SR: Qualitative comparison of sentiment predictions. The first three samples in the evaluation dataset are presented. Each example includes the transcript, ground-truth sentiment label, and model predictions with their inferred sentiments. Sentiment labels are in blue, and quotes from the speech content are in orange. Interpretive and evidence-grounded reasoning better align with the ground truth, offering richer justifications and improved sentiment recognition.