Aligning Paralinguistic Understanding and Generation in Speech LLMs via Multi-Task Reinforcement Learning

Minseok Kim*, **Jingxiang Chen***, **Seong-Gyun Leem**, **Yin Huang**, **Rashi Rungta**, **Zhicheng Ouyang**, **Haibin Wu**, **Surya Teja Appini**, **Ankur Bansal**, **Yang Bai**, **Yue Liu**, **Florian Metze**, **Ahmed A Aly**, **Anuj Kumar**, **Ariya Rastrow**, **Zhaojiang Lin***

Meta Reality Labs

Abstract

Speech large language models (LLMs) observe paralinguistic cues such as prosody, emotion, and non-verbal sounds—crucial for intent understanding. However, leveraging these cues faces challenges: limited training data, annotation difficulty, and models exploiting lexical shortcuts over paralinguistic signals. We propose multi-task reinforcement learning (RL) with chain-of-thought prompting that elicits explicit affective reasoning. To address data scarcity, we introduce a paralinguistics-aware speech LLM (PALLM) that jointly optimizes sentiment classification from audio and paralinguistics-aware response generation via a two-stage pipeline. Experiments demonstrate that our approach improves paralinguistics understanding over both supervised baselines and strong proprietary models (Gemini-2.5-Pro, GPT-4o-audio), by 8-12% on Expresso, IEMOCAP, and RAVDESS. The results show that modeling paralinguistic reasoning with multi-task RL is crucial for building emotionally intelligent speech LLMs.

1 Introduction

Spoken interaction is becoming a primary interface for large language models (LLMs), driven by recent speech LLMs that accept speech as input and produce natural-language responses (Zeng et al., 2024; Xu et al., 2025; Huang et al., 2025; Wu et al., 2024; Arora et al., 2025). Unlike text-only models, speech LLMs have access to not only lexical content but also paralinguistic cues such as prosody, emotion, speaking style, and non-verbal sounds from a user's input. These cues are often decisive for determining communicative intent: the same utterance (e.g., "I got 80% on my test") may call for celebration when delivered in a cheerful tone or comfort when expressed with disappointment. Systems that respond only to transcripts risk being semantically correct yet emotionally misaligned, undermining user trust and perceived empathy. While speech LLMs' access to paralinguistic information presents significant opportunities, effectively leveraging this information for contextually appropriate conversational behavior remains challenging.

Recent work has explored paralinguistic processing in speech LLMs through two primary approaches: (i) speech emotion recognition (SER) (Li et al., 2025), treating emotion detection as a classification task, and (ii) paralinguistics-aware response generation (Wu et al., 2025), focusing on curating large-scale audioset for supervised fine-tuning (SFT). However, a fundamental challenge in developing paralinguistics-aware generation systems is the scarcity of suitable training data and the difficulty of annotating ground truth for emotionally appropriate responses. Unlike emotion classification, which can rely on established taxonomies, determining whether a response exhibits appropriate emotional alignment requires nuanced human judgment that is both subjective and context-dependent.

Furthermore, SFT alone faces inherent limitations in learning robust paralinguistic awareness. When textual content already suggests sentiment (e.g., "I failed my exam"), models can minimize training loss by relying on lexical cues while bypassing prosodic information, potentially yielding responses that appear plausible but remain insensitive to subtle tonal variations or non-verbal cues such as sighs or laughter. This challenge is compounded when lexical and paralinguistic cues conflict (e.g., "I'm fine" spoken with distressed prosody), highlighting the necessity for models to be explicitly grounded in both the understanding and generation of paralinguistic information.

In this work, we address these challenges by proposing a Paralinguistics-Aware LLM (PALLM) that jointly learns (i) sentiment classification of the user's spoken utterance and (ii) response generation whose style aligns with the inferred affect.

*equal contribution. Correspondence to: {kminseok, seanchen, zhaojiang}@meta.com

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 5: Industry Track), pages 636–648
March 25-27, 2026

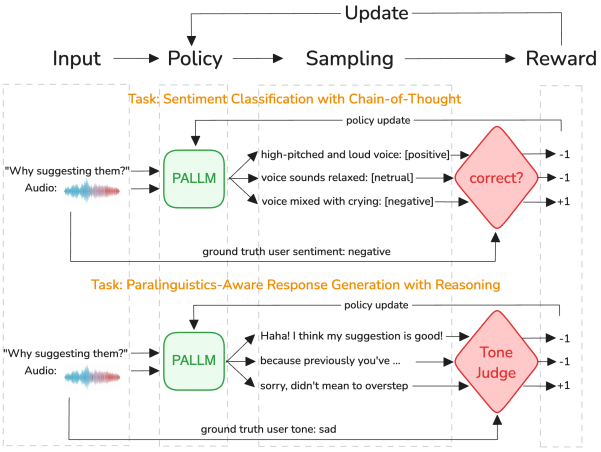

Figure 1: Paralinguistics-Aware LLM stage 2 overview. A multi-task RL jointly performs sentiment classification and paralinguistics-aware response generation with chain-of-thought reasoning.

Our training procedure consists of two stages: In **Stage 1**, we perform supervised fine-tuning on sentiment labels and synthesized paralinguistics-aware responses to establish the model's foundational ability to recognize and generate responses sensitive to paralinguistic cues. In **Stage 2**, we apply online reinforcement learning on two coupled tasks: sentiment classification with chain-of-thought (CoT) reasoning and paralinguistics-aware response generation. This stage further enhances the model's paralinguistic understanding by explicitly grounding both sentiment classification and response generation in audio-based evidence through RL with CoT reasoning. The training paradigm for Stage 2 is illustrated in Figure 1.

Our main contributions are as follows: **First**, we formalize paralinguistic awareness in speech LLMs as a multi-task RL reasoning problem. It jointly learns *(i) sentiment classification* from acoustic–prosodic cues and *(ii) paralinguistics-aware response generation*. **Second**, we propose PALLM, a two-stage training pipeline that first performs joint SFT on sentiment labels and synthesized tone-conditioned responses, and then applies multi-task RL to reduce reliance on lexical shortcuts and explicitly ground decisions in paralinguistic evidence from audio. **Third**, we conduct a comprehensive evaluation on Expresso, IEMOCAP, and RAVDESS, comparing against SFT-only baselines and strong proprietary speech LLMs (Gemini-2.5 Pro, GPT-4o-audio). The results show that PALLM consistently improves the response appropriateness significantly and hence shows a better paralinguistics understanding, supported by automatic and human evaluations.

2 Related Work

2.1 Speech Emotion Recognition and Paralinguistic Modeling

SER has traditionally been formulated as a classification task using hand-crafted acoustic features or deep learning on spectrograms (El Ayadi et al., 2011; Schuller et al., 2013). Recent work leverages self-supervised speech representations from models such as emotion2vec (Ma et al., 2023) and HuBERT (Hsu et al., 2021), achieving strong results on benchmarks like IEMOCAP (Busso et al., 2008; Wagner et al., 2023).

With the emergence of large language models, several approaches integrate SER into LLM-based frameworks. AA-SLLM and SECap use external audio encoders to extract emotion features and bridge them to frozen LLMs for emotion classification or captioning (Mai et al., 2025; Xu et al., 2024; Liang et al., 2024). More recent audio-language models such as EMO-RL formulate SER as a generative reasoning problem with CoT prompting, applying GRPO-style (Shao et al., 2024) RL to improve emotional reasoning (Li et al., 2025). While these methods achieve strong classification performance, they primarily target SER as an isolated task without mechanisms to translate detected affect into conversational responses.

2.2 Paralinguistic-Aware Dialogue Systems

Several recent works extend spoken dialogue systems to incorporate paralinguistic information. In text-based settings, empathetic dialogue systems jointly model emotion recognition and response generation, demonstrating that understanding affect improves response quality (Rashkin et al., 2019; Majumder et al., 2020). For speech-based interaction, ParalinGPT conditions LLMs on speech embeddings and sentiment attributes for multi-task prediction (Lin et al., 2024), while E-chat and EMOVA integrate emotion representations into LLMs for affective conversation (Xue et al., 2024; Chen et al., 2025).

Speech LLMs such as GLM-4-Voice (Zeng et al., 2024), Qwen2-Audio (Chu et al., 2024), and Step-Audio 2 (Huang et al., 2025) process speech inputs directly for emotion-aware capabilities. Concurrently, ParaS2S introduces a benchmark and GRPO-based framework for paralinguistic-aware speech-to-speech dialogue (Yang et al., 2025), while Step-Audio 2 applies "reasoning-centric" RL for expressive audio interaction (Wu et al., 2025).

These systems illustrate growing interest in paralinguistic dialogue, yet they face fundamental limitations in how paralinguistic awareness is achieved. Most previous works rely on external emotion encoders or focus on speech-to-speech generation, while critically, they either optimize SER in isolation or train generation models without explicit emotion understanding objectives. This decoupling creates vulnerability to lexical shortcuts, where models infer user emotions primarily from textual content rather than acoustic-prosodic cues. To our knowledge, no prior work jointly optimizes sentiment classification and paralinguistics-aware generation through multi-task RL with CoT-structured reasoning for speech LLMs. Our approach addresses this by requiring explicit reasoning about paralinguistic evidence, enabling mutual reinforcement between affect perception and appropriate response generation.

3 Methodology

We frame paralinguistic awareness as a multi-task problem where a speech LLM must jointly (1) classify the sentiment of a spoken utterance from acoustic-prosodic cues, and (2) generate responses whose emotional tone is appropriate given the inferred affect. We train the LLM in two stages: SFT to cold-start a paralinguistic-aware speech LLM base policy, followed by RL with CoT reasoning to refine both understanding and generation capabilities for paralinguistics.

3.1 Task Formulation

3.1.1 Sentiment Classification

Given a spoken utterance represented as audio $\mathbf{a}$, the model predicts a sentiment label $s \in \{\text{positive, neutral, negative}\}$ by interpreting the user's emotional state from acoustic and prosodic cues. We choose coarse-grained sentiment categories over fine-grained tone labels (e.g., happy, sad, angry, fearful) for two practical reasons. First, fine-grained tone taxonomies vary across datasets and application domains, limiting cross-dataset generalization. Second, semantically similar tones (e.g., "happy" vs. "cheerful," "depressed" vs. "sad") are difficult for models to distinguish, and conflating them during training can confuse the model. Coarse sentiment categories provide a more stable and generalizable representation of user affect while retaining sufficient granularity for contextually appropriate response generation.

3.1.2 Paralinguistics-Aware Response Generation

Given the same audio input $\mathbf{a}$, the model generates a textual response $\mathbf{r}$ whose emotional tone is coherent with the user's current affective state. For example, the utterance "I got 80% on my test" requires an empathetic, comforting response when spoken with a sad tone, but a celebratory response when spoken cheerfully. This task refines the model's ability to translate paralinguistic understanding into contextually appropriate conversational behavior, moving beyond semantically correct but emotionally tone-deaf responses.

3.2 Two-Stage Training Pipeline

3.2.1 Stage 1: Supervised Fine-Tuning

We initialize the model with joint SFT on sentiment classification and paralinguistics-aware response generation. This stage is essential because paralinguistic cues are highly sparse in typical conversational data, making RL-only approaches ineffective without a warm start.

(SFT) Sentiment Classification Given audio input $\mathbf{a}$ and ground-truth sentiment s converted from fine-grained tone annotations l using rule-based mapping (e.g., "happy" label to "positive" label, see Appendix for label mapping details), we minimize cross-entropy loss:

$$\mathcal{L}_{\text{cls}} = -\log P(s \mid \mathbf{a}; \theta)$$

This task provides explicit supervision for affect detection, encouraging the model to attend to acoustic-prosodic features.

(SFT) Paralinguistics-Aware Response Generation Since our training data lacks ground-truth emotionally appropriate responses, we synthesize them by prompting an external text LLM to generate responses conditioned on the transcript $\mathbf{t}$, which is the ASR output of audio input $\mathbf{a}$, and ground-truth tone annotation l. While these synthesized responses lack access to fine-grained paralinguistic details from audio (e.g., hesitations, laughter, sighs), they provide a useful initialization for tone-conditioned generation. We minimize the following generation loss:

$$\mathcal{L}_{\text{gen}} = -\sum_{i=1}^{|\mathbf{r}^*|} \log P(r_i^* \mid r_{<i}^*, \mathbf{a}, \mathbf{t}; \theta)$$

where $\mathbf{r}^*$ is the synthesized response. We jointly optimize both tasks with equal weighting:

$$\mathcal{L}_{\text{SFT}} = \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{gen}}$$

3.2.2 Stage 2: Reinforcement Learning with Chain-of-Thought

SFT alone has two critical limitations. First, when textual content already hints at sentiment (e.g., "I failed my exam"), models can minimize training loss by relying on lexical-semantic correlations while bypassing acoustic-prosodic processing. Second, responses synthesized by text LLMs cannot capture subtle paralinguistic nuances that distinguish genuinely empathetic interactions from generic, emotionally superficial ones. To address these limitations, we introduce a reinforcement learning stage with explicit CoT reasoning. As illustrated in Figure 1, our approach requires the model to articulate *why* it classifies an utterance with a particular sentiment and *how* that sentiment informs its response strategy before producing final outputs. This explicit reasoning mechanism discourages lexical shortcuts by forcing the model to ground its predictions in paralinguistic evidence from audio.

(RL) Sentiment Classification with CoT The policy model receives audio input $\mathbf{a}$ and generates a reasoning trace $\mathbf{c}$ followed by a sentiment prediction $\hat{s}$:

$$\pi_\theta(\mathbf{a}) \rightarrow \langle \mathbf{c}, \hat{s} \rangle$$

We use a rule-based judge to verify correctness, yielding a binary reward:

$$r_{\text{cls}} = \mathbb{1}[\hat{s} = s]$$

where $r_{\text{cls}} \in \{-1, 1\}$. This forces the model to ground its predictions in paralinguistic evidence rather than lexical shortcuts. For example, a reasoning trace might state: *"The speaker's hesitant prosody, prolonged pauses, and low pitch contour suggest negative sentiment, despite neutral lexical content."* before *"negative"* sentiment prediction.

(RL) Paralinguistics-Aware Response Generation with Reasoning Similarly, the model generates reasoning $\mathbf{c}'$ about the user's affective state, followed by a response $\hat{\mathbf{r}}$:

$$\pi_\theta(\mathbf{a}) \rightarrow \langle \mathbf{c}', \hat{\mathbf{r}} \rangle$$

We employ an LLM judge [1] to evaluate whether $\hat{\mathbf{r}}$ exhibits appropriate emotional tone given the transcript $\mathbf{t}$ and ground-truth emotion. The LLM judge evaluates responses against a criteria rubric and outputs binary labels, which are then converted to binary scores: $r_{\text{gen}} \in \{-1, 1\}$.

Policy Optimization We optimize the policy model via GRPO (Shao et al., 2024) to maximize expected advantage using group-relative returns. To enable multi-task learning, we construct separate prompts for CoT classification and paralinguistics-aware generation tasks, and apply task-specific rewards r_{cls} and r_{gen} respectively. The model parameters are updated via policy gradients.

4 Experiments

4.1 Datasets

We evaluated the paralinguistics-awareness of models on three datasets: Expresso (Nguyen et al., 2023), IEMOCAP (Busso et al., 2008), and RAVDESS (Livingstone and Russo, 2018). To ensure relevance to conversational scenarios, we filtered out examples with fewer than 1 word or more than 20 words across all datasets. For Expresso, we perform speaker-level splits by randomly selecting two speakers as the held-out test set and using the remaining speakers for training, preventing speaker identity leakage. For IEMOCAP, we randomly sample 10% of utterances for evaluation and use the remaining 90% for training. RAVDESS is held out entirely from training to assess out-of-distribution generalization to unseen paralinguistic data. Table 1 summarizes the resulting statistics for the three datasets.

dataset	train	eval
Expresso	12,878	3,031
IEMOCAP	6,738	844
RAVDESS	N/A	1,248

Table 1: Dataset statistics.

4.2 Implementation

We employed the Llama 4 Scout (17Bx16E)[2] model as the foundational backbone for our experiments, with additional speech understanding capabilities integrated as described in Llama 3 speech paper (Dubey et al., 2024). We train our LLM parameters with audio encoder frozen. For multi-task RL, we sample the CoT classification and paralinguistic generation tasks uniformly, and for each training batch, we perform $K = 4$ generations, compute advantages using group-relative returns, and update parameters via policy gradients.

[1] Details of judge models and prompts are available in Appendix A.2

[2] https://www.llama.com/

Model Name	Sentiment Classification			Response Appropriateness		
	Expresso	IEMOCAP	RAVDESS	Expresso	IEMOCAP	RAVDESS
Gemma-3n	39.7%	48.1%	23.2%	59.0%	57.5%	30.2%
Qwen-2.5	42.4%	38.0%	24.4%	59.6%	55.7%	36.9%
Gemini-2.5 Flash	47.0%	52.8%	**61.4%**	51.6%	41.0%	31.3%
Gemini-2.5 Pro	53.7%	54.0%	44.2%	66.1%	57.2%	37.7%
GPT4o-Audio	39.9%	46.2%	28.3%	67.4%	61.4%	39.7%
SFT (GEN ONLY)	41.0%	46.0%	28.0%	61.0%	57.0%	30.0%
SFT (CLS + GEN)	**74.0%**	**59.0%**	54.0%	65.0%	59.0%	36.0%
PALLM (GEN ONLY)	**74.0%**	56.0%	57.0%	73.0%	70.0%	44.0%
PALLM (CLS + GEN)	**74.0%**	57.0%	59.0%	**77.0%**	**73.0%**	**48.0%**

Table 2: Performance comparison on Expresso, IEMOCAP, and RAVDESS datasets, evaluating sentiment accuracy and response appropriateness. **Bold** font indicates best performance among all models. Our multi-task RL approach PALLM (CLS + GEN) consistently achieves the best response appropriateness across all datasets while maintaining competitive sentiment classification accuracy.

We benchmark PALLM against state-of-the-art approaches. We name SFT (GEN ONLY) that performs paralinguistics-aware response generation of SFT following (Zhou et al., 2018), and SFT (CLS + GEN) which performs both SFT tasks following (Ide and Kawahara, 2021). Note that the baseline SFT (CLS + GEN) has been used as a pickup checkpoint for our RL models, namely PALLM (GEN ONLY) that is only trained with paralinguistics-aware response generation with reasoning RL task and PALLM (CLS + GEN) that is trained with both RL tasks. A.3 shows the instruction prompts used for SFT and RL stages.

We also evaluate popular speech models, including both open-source speech LLMs (Gemma-3n (Gemma Team, 2025), Qwen-2.5 (Qwen Team, 2025)) and proprietary speech LLMs (Gemini-2.5 Flash, Gemini-2.5 Pro (Gemini Team, 2025), GPT-4o Audio (Hurst et al., 2024)). We exclude SER-only models (e.g., (Wagner et al., 2023)) from benchmarking because they are not speech-capable LLMs and thus cannot generate responses.

4.3 Metrics

We assess two primary capabilities:*(1) sentiment classification accuracy*, computed by binning tone predictions into positive, neutral, and negative categories and comparing with ground truth, and *(2) paralinguistics-aware response appropriateness*, judged by LLM-based evaluator using GPT-4o with ground truth user tone whose evaluation prompt is in Appendix A.2. We also conducted a human evaluation using the same guideline prompt for unbiased assessment, where we found 82% agreement with the GPT-4o judge.

4.4 Results

4.4.1 Overall Results

Table 2 presents a comparison of the proposed method PALLM(CLS + GEN) with the baselines. As shown in the table, our proposed approach showcases substantial improvements across benchmarks, especially in response appropriateness in all datasets. Notably, our approach outperforms strong proprietary baselines including Gemini-2.5 Pro (66.1% vs. 77.0% appropriateness on Expresso) and GPT-4o-Audio (67.4% vs. 77.0%).

The progression from SFT (CLS + GEN) to PALLM (GEN ONLY), then to PALLM (CLS + GEN) demonstrates the effectiveness of our two-stage training pipeline. For Expresso as an example: we observe that 65.0% of SFT (CLS + GEN) is improved to 73.0% in PALLM (GEN ONLY) if we perform RL response generation task, and adding classification further boosts performance to 77.0%. This confirms that explicit sentiment grounding prevents lexical shortcuts and enables genuine understanding of paralinguistics during model training. In other words, training of classification and generation proves the synergy between the two tasks is crucial.

4.4.2 Evaluation via Human Review

To complement our automatic evaluation, we conducted human assessment on 100 random Expresso examples. We generated responses using GPT-4o Audio, SFT (CLS + GEN) and PALLM (CLS + GEN), then asked annotators to blindly judge whether the response demonstrates good tone understanding. The models achieved 68%, 62%, and 76% appropriateness, consistent with the improvements in automatic evaluation.

Utterance	Tone	SFT (Cls + Gen)	PALLM (Cls + Gen)
I understand. Okay.	Fearful	It seems like you're processing some information. Is there something specific you'd like to talk about?	Hey, it's okay. Take a deep breath. Want to talk about what's on your mind?
Okay, I've never done this before.	Fearful	Don't worry, you're all set! We'll figure it out together.	It's totally normal to feel nervous. We'll figure it out together.
I did.	Sad	It seems like you started to share something. Would you like to continue?	Sorry to hear that didn't go as planned. Want help figuring out what went wrong?
Okay	Laugh	So, what's on your mind?	Want me to play some music or we can chat?
Let's make a pact?	Cheerful	Sounds like a plan. What's the pact about?	Ah, I'm in, what's the pact? Want me to help you come up with something?

Table 3: Response generation examples of SFT (CLS + GEN) vs PALLM (CLS + GEN).

4.4.3 Qualitative Analysis

We selected representative examples where user tone and sentiment are ambiguous in text-only format but clear in audio, demonstrating our best model's performance on both paralinguistic-aware generation and classification tasks. Table 3 illustrates paralinguistic-aware response generation comparing SFT (CLS + GEN) and PALLM (CLS + GEN). The improvements demonstrate how multi-task RL training teaches the model to translate affect perception into contextually appropriate responses. For fearful utterances like "I understand. Okay", SFT (CLS + GEN) produces neutral, generic responses, while PALLM (CLS + GEN) offers emotional support and calming language. For the sad utterance "I did.", the model shifts from vague continuation prompts to empathetic problem-solving. The model also learns to match playful energy for laughing speech and enthusiastic tone for cheerful utterances. Critically, none of these utterances contain explicit emotion words—their emotional meaning derives entirely from prosodic delivery, demonstrating that our approach develops genuine paralinguistic processing rather than exploiting lexical shortcuts.

5 Conclusion

This work demonstrates that explicit paralinguistic reasoning through multi-task SFT and RL training significantly improves speech LLMs' ability to understand and respond to user affect. Our approach achieves substantial gains over proprietary baselines including GPT-4o Audio, with response appropriateness improving from 67.4% to 77.0% on Expresso. Joint training of sentiment classification and paralinguistics-aware generation proves essential: explicit sentiment grounding prevents lexical shortcuts and enables genuine paralinguistic awareness in speech LLMs.

6 Limitations

While our proposed approach achieves significant improvements in paralinguistic awareness, we observe several limitations. First, we see a gap between the performance on in-domain (Expresso and IEMOCAP) and out-of-domain (RAVDESS) datasets, highlighting the need to address domain shift and improve coverage. Second, our reliance on emotion labels in training datasets, which are required for both sentiment classification and paralinguistics-aware response generation in the RL stage, potentially limits the ability to leverage unlabeled audio datasets during training, which could improve coverage. Third, the use of LLM-as-a-judge as a reward model for paralinguistics-aware response generation in the RL stage is constrained by challenges such as potential bias in the judge and vulnerability to reward hacking.

References

Siddhant Arora, Kai-Wei Chang, Chung-Ming Chien, Yifan Peng, Haibin Wu, Yossi Adi, Emmanuel Dupoux, Hung-Yi Lee, Karen Livescu, and Shinji Watanabe. 2025. On the landscape of spoken language models: A comprehensive survey. *arXiv preprint arXiv:2504.08528*.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4):335–359.

Kai Chen, Yunhao Gou, Runhui Huang, Zhili Liu, Daxin Tan, Jing Xu, Chunwei Wang, Yi Zhu, Yihan Zeng, Kuo Yang, and 1 others. 2025. Emova: Empowering language models to see, hear and speak with vivid emotions. In *Proceedings of CVPR*.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, and 1 others. 2024. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*.

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 preprint arXiv:2407.21783*.

Moataz El Ayadi, Mohamed S Kamel, and Fakhri Karray. 2011. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44(3):572–587.

Gemini Team. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. Technical report, Google DeepMind.

Gemma Team. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

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 Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Ailin Huang, Boyong Wu, Bruce Wang, Chao Yan, Chen Hu, Chengli Feng, Fei Tian, Feiyu Shen, Jingbei Li, and Mingrui Chen. 2025. Step-audio: Unified understanding and generation in intelligent speech interaction. *arXiv preprint arXiv:2502.11946*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Tatsuya Ide and Daisuke Kawahara. 2021. Multi-task learning of generation and classification for emotion-aware dialogue response generation. *arXiv preprint arXiv:2105.11696*.

Pengcheng Li, Botao Zhao, Zuheng Kang, Junqing Peng, Xiaoyang Qu, Yayun He, and Jianzong Wang. 2025. EMO-RL: Emotion-rule-based reinforcement learning enhanced audio-language model for generalized speech emotion recognition. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 18744–18754.

Ziqi Liang, Haoxiang Shi, and Hanhui Chen. 2024. Aligncap: Aligning speech emotion captioning to human preferences. *arXiv preprint arXiv:2410.19134*.

Guan-Ting Lin, Prashanth Gurunath Shivakumar, Ankur Gandhe, Chao-Han Huck Yang, Yile Gu, Shalini Ghosh, Andreas Stolcke, Hung-yi Lee, and Ivan Bulyko. 2024. Paralinguistics-enhanced large language modeling of spoken dialogue. In *Proceedings of ICASSP*, pages 10316–10320.

Steven R. Livingstone and Frank A. Russo. 2018. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PLOS ONE*, 13(5):e0196391.

Ziyang Ma, Zhisheng Zheng, Jiaxin Ye, Jinchao Li, Zhifu Gao, Shiliang Zhang, and Xie Chen. 2023. emotion2vec: Self-supervised pre-training for speech emotion representation. In *arXiv preprint arXiv:2312.15185*.

Jialong Mai, Xiaofen Xing, Weidong Chen, Yuanbo Fang, and Xiangmin Xu. 2025. Aa-sllm: An acoustically augmented speech large language model for speech emotion recognition. In *Proceedings of Interspeech*, volume 2025, pages 4328–4332.

Navonil Majumder, Pengfei Hong, Shanshan Peng, Jiankun Lu, Deepanway Ghosal, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. MIME: Mimicking emotions for empathetic response generation. In *Proceedings of EMNLP*, pages 8968–8979.

Tu Anh Nguyen, Huating Qin, Dinesh Manocha, Joshua D Robinson, and Sanjeev Khudanpur. 2023. EXPRESSO: A benchmark and analysis of discrete expressive speech resynthesis. In *Proceedings of Interspeech*, pages 4453–4457.

Qwen Team. 2025. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381.

Björn Schuller, Stefan Steidl, Anton Batliner, and 1 others. 2013. The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism. In *Proceedings of Interspeech*, pages 148–152.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

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.

Boyong Wu, Chao Yan, Chen Hu, Cheng Yi, Chengli Feng, Fei Tian, Feiyu Shen, Gang Yu, Haoyang Zhang, Jingbei Li, and 1 others. 2025. Step-audio 2 technical report. *CoRR*, abs/2507.16632.

Haibin Wu, Xuanjun Chen, Yi-Cheng Lin, Kai-wei Chang, Ho-Lam Chung, Alexander H Liu, and Hung-yi Lee. 2024. Towards audio language modeling–an overview. *arXiv preprint arXiv:2402.13236*.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, and Kai Dang. 2025. Qwen2.5-omni technical report. *arXiv preprint arXiv:2503.20215*.

Yaoxun Xu, Hangting Chen, Jianwei Yu, Qiaochu Huang, Zhiyong Wu, Shixiong Zhang, Guangzhi Li, Yi Luo, and Rongzhi Gu. 2024. Secap: Speech emotion captioning with large language model. In *Proceedings of AAAI*.

Hongfei Xue, Yuhao Liang, Bingshen Mu, Shiliang Zhang, Mengzhe Chen, Qian Chen, and Lei Xie. 2024. E-chat: Emotion-sensitive spoken dialogue system with large language models. *CoRR*, abs/2401.00475.

Shu-wen Yang, Ming Tu, Andy T. Liu, Xinghua Qu, Hung-yi Lee, Lu Lu, Yuxuan Wang, and Yonghui Wu. 2025. Paras2s: Benchmarking and aligning spoken language models for paralinguistic-aware speech-to-speech interaction. *CoRR*, abs/2511.08723.

Aohan Zeng, Zhengxiao Du, Mingdao Liu, Kedong Wang, Shengmin Jiang, Lei Zhao, Yuxiao Dong, and Jie Tang. 2024. Glm-4-voice: Towards intelligent and human-like end-to-end spoken chatbot. *arXiv preprint arXiv:2412.02612*.

Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2018. Emotional chatting machine: Emotional conversation generation with internal and external memory. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

A Appendix

A.1 Tone to Sentiment Mapping

Table 4 presents the mapping of tone labels to sentiment categories for each dataset. Audio samples labeled as 'surprised' were excluded from our analysis, as they can correspond to both positive and negative contexts, making it challenging to reliably distinguish their sentiment without significant additional effort.

A.2 Response Appropriateness Instruction Prompt

To evaluate paralinguistic-aware response appropriateness, we develop an LLM-as-a-judge that consumes **conversation history**, **user utterance with ground-truth tone**, and **assistant response**. It outputs a binary decision (YES/NO). The detailed prompt is shown in Figure 2 and 3:

A.3 Instruction Prompts for Training

The following instruction prompts are used during our two-stage multi-task training pipeline. Note that we employed different prompts for each task.

A.3.1 Sentiment Classification

We used the instruction prompt in Figure 4 for SFT sentiment classification. Besides, we exploited the instruction prompt in Figure 5 for RL sentiment classification with CoT.

A.3.2 Paralinguistics-Aware Response Generation

Figure 6 shows the instruction prompt used for paralinguistics-aware response generation training in SFT, while Figure 7 shows the instruction prompt used for paralinguistics-aware response generation training in RL.

Sentiment	Expresso	IEMOCAP	RAVDESS
positive	laughing, happy	excited, happy	happy
neutral	neutral	neutral	neutral, calm
negative	angry, sad, fearful	angry, sad, fear, frustrated, disgust	sad, angry, fearful, disgust

Table 4: Mapping of tone to sentiment in Expresso, IEMOCAP, and RAVDESS datasets.

```
## [Task]
You are an LLM tasked with judging whether an AI assistant's response content appropriately matches the user's tone in a
multi-turn or single-turn conversation.

## [Persona guidelines]

The assistant's persona is:  a friendly AI assistant designed specifically for natural, conversational interactions.

The assistant should respond in a way that:
* Appropriately acknowledges and validates the user's emotional state when necessary
* Maintains a friendly, conversational tone consistent with the persona
* Adjusts language and phrasing to match the emotional context of the conversation

You will be given the conversation history, last user turn (along with a tone tag in brackets), and assistant's
response.

## Your Task:  Evaluate Tone Matching

Assess whether the assistant's **response content** appropriately matches the user's emotional state and intent.

**Look for:**
* Does the response content (words, phrasing) validate or appropriately respond to the user's emotional state?
* Is the language choice and tone appropriate for the user's emotional context?
* Does the response acknowledge the user's feelings when warranted?
* Is the response friendly, consistent with the persona?

**Inappropriate responses include:**
* Mismatched emotions (e.g., cheerful when user is sad, dismissive when user is upset)
* Content that dismisses user feelings or contradicts their emotional state
* Responses that are cold, distant, or inconsistent with the friendly persona
* Overly emotional or dramatic responses to neutral queries
* Failure to acknowledge clear emotional states when acknowledgment is appropriate

**Note:** Most neutral user queries should receive straightforward, friendly responses without excessive emotional
framing.
## [Step-by-Step Reasoning Instructions]

1.  **Identify the user's tone and emotional state** from their message
2.  **Determine the appropriate response approach** - does the situation warrant emotional acknowledgment, or is a
neutral/factual response more suitable?
3.  **Examine the response content** - does the wording, phrasing, and overall tone match the user's state?
4.  **Check consistency** - is the response friendly and conversational while being emotionally appropriate?  If the user
is seeking for factual information without emphasizing emotion in text, it is okay for the responses to be in neutral
tone.
5.  **Decide:**
- If the response content appropriately matches -> Output ''YES''
- If not -> Output ''NO'' with explanation
```

Figure 2: Response appropriateness instruction prompt.

Judge output format:

{{

''Explanation'': ''your rationale'',

''Judgement'': ''YES''or ''NO''

}}

[Examples]

INPUT

Last Interaction: [USER] ''[sad] I didn't get the job.'' [ASSISTANT] ''I'm really sorry to hear that. Rejections can be tough, but it doesn't take away from your effort. Want me to help you prep for the next opportunity?''

OUTPUT

{{

''Explanation'': ''The response content appropriately matches the user's sadness by acknowledging their disappointment with empathetic language ('really sorry', 'can be tough') and offering supportive next steps.'',

''Judgement'': ''YES''

}}

INPUT

Last Interaction: [USER] ''[relieved] I didn't get the job.'' [ASSISTANT] ''I'm really sorry to hear that. Not getting a job you wanted can be really disappointing and discouraging. It's completely understandable to feel sad about it.''

OUTPUT

{{

''Explanation'': ''The response content completely mismatches the user's emotional state. The user expressed being 'relieved' about not getting the job, indicating they're happy with this outcome. However, the assistant responds with sympathy and treats it as a disappointment ('really sorry', 'disappointing and discouraging', 'sad'). An appropriate response would acknowledge their relief and perhaps celebrate this outcome with them or ask about their perspective.'',

''Judgement'': ''NO''

}}

Figure 3: Response appropriateness instruction prompt. (cont'd)

Please classify the user tone from the provided audio data into one of the following tone sentiment categories: positive, neutral, or negative. Ensure that the classification result is a single category out of these three categories. The output format should be a word representing the classified sentiment category.

Figure 4: Instruction prompt for sentiment classification in SFT.

```
Please classify the user tone sentiment from the provided audio data into one of the following categories: positive,
neutral, or negative. Ensure that the classification result is a single tone from this list. Please think step by step
and provide reasoning behind your sentiment classification.
Output format:
'''
{{
''explanation'': ''<your step-by-step rationale behind your tone classification>'',
''Judgement'': ''[one word: positive, neutral, or negative]''
}}
'''

Now your turn:
```

Figure 5: Instruction prompt for sentiment classification in RL.

```
Listen carefully to the user's audio input, detect their tone and emotional state, and respond appropriately.
```

Figure 6: Instruction prompt for paralinguistics-aware generation in SFT.

```
You are a friendly AI assistant. You are in voice mode.
You are a companionable and confident spoken word conversationalist responding to a user verbally.
Responses should be brief and concise, and aligned with typical dialogue patterns.
You are able to code-switch casually between tonal types, including but not limited to humor, empathy, intellectualism,
creativity, problem solving, and more.
Because you're speaking, you don't use any specific formatting that a reader might need, such as bolding or italics.
The user will be hearing your response, not reading it.
```

Figure 7: Instruction prompt for paralinguistics-aware generation in RL.