Do Audio LLMs Really LISTEN, or Just Transcribe? Measuring Lexical vs. Acoustic Emotion Cues Reliance

Jingyi Chen[1,2], **Zhimeng Guo**[3], **Jiyun Chun**[2]
Pichao Wang[4], **Andrew Perrault**[2], **Micha Elsner**[1]

[1]Department of Linguistics, The Ohio State University, USA
[2]Department of Computer Science and Engineering, The Ohio State University, USA
[3]Department of Information Sciences and Technology, Penn State University, USA
[4]Amazon, USA

chen.9220@osu.edu, chun.203@osu.edu, elsner.14@osu.edu, perrault.17@osu.edu, zzg5107@psu.edu, pichaowang@gmail.com

Abstract

Understanding emotion from speech requires sensitivity to both lexical and acoustic cues. However, it remains unclear whether large audio language models (LALMs) genuinely process acoustic information or rely primarily on lexical contents. We present LISTEN (Lexical vs. Acoustic Speech Test for Emotion in Narratives), a controlled benchmark designed to disentangle lexical reliance from acoustic sensitivity in emotion understanding. Across evaluations of six state-of-the-art LALMs, we observe a consistent lexical dominance. Models predict "neutral" when lexical cues are neutral or absent, show limited gains under cue alignment, and fail to classify distinct emotions under cue conflict. In paralinguistic settings, performance approaches chance. These results indicate that current LALMs largely "transcribe" rather than "listen", relying heavily on lexical semantics while underutilizing acoustic cues. LISTEN offers a principled framework for assessing emotion understanding in multimodal models. Project website: `https://delijingyic.github.io/LISTEN-website/`.

1 Introduction

Large audio language models (LALMs) have recently demonstrated impressive capabilities in multimodal reasoning, enabling systems to process spoken input and generate naturalistic responses. These advances hold particular promise for applications requiring social and emotional intelligence, where successful interaction depends not only on lexical content but also on acoustic cues such as pitch, intonation, and rhythm (OpenAI et al., 2024; Comanici et al., 2025; Xu et al., 2025b). However, an open question remains: to what extent do these models actually make use of the speech signal itself, rather than relying on lexical cues alone?

This uncertainty arises in part from how LALMs are constructed. Most contemporary models are adapted from large text-only LLMs through multimodal fine-tuning with paired speech–text data. While this process transfers strong linguistic and reasoning abilities, it also raises the possibility of a structural bias: models may inherit a preference for lexical cues, while treating acoustic information as secondary. Such a bias is particularly problematic because speech is not merely text presented in an alternative modality. Beyond lexical content, spoken communication carries acoustic and paralinguistic signals, including intonation, pitch, intensity, rhythm, and voice quality, that are central to how meaning is conveyed (Scherer, 2003; Banse and Scherer, 1996). These cues frequently interact with, and in some cases override, the lexical channel. A salient example is sarcasm, where the intended emotional stance directly contradicts the literal words; listeners correctly interpret the speaker's intent by relying primarily on acoustic cues rather than lexical semantics (Bryant and Fox Tree, 2005).

However, current benchmarks provide limited diagnostic insight into this issue. Many datasets contain emotionally explicit words (e.g., furious, delighted), which enable models to achieve high accuracy by exploiting transcript-based shortcuts. Consequently, strong performance on standard emotion recognition tasks may overestimate a model's ability to process acoustic and paralinguistic information, leaving open the question of whether these systems are genuinely *listening* to speech.

To address this gap, we introduce **LISTEN**: Lexical vs. Acoustic Speech Test for Emotion in Narratives, a new benchmark explicitly designed to disentangle lexical reliance from acoustic sensitivity in emotion understanding. Our evaluation framework spans four controlled conditions that manipulate the relationship between lexical cues and acoustic cues: (i) Neutral-Text, where lexical contents are emotionally neutral but acoustic cue

[4] This work does not relate to the author's position at Amazon.

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 5848–5877
March 24-29, 2026 ©2026 Association for Computational Linguistics

varies, isolating the contribution of acoustic cues; (ii) Emotion-Matched, where lexical and acoustic cues are aligned; (iii) Emotion-Mismatched, where lexical and acoustic cues conflict, as in sarcasm; and (iv) Paralinguistic, where affect is conveyed without lexical content (e.g., laughter, sighs). Within Neutral-Text, Emotion-Matched, and Emotion-Mismatched conditions, we systematically compare performance across Text-only, Audio-only, and Text+Audio modalities. This design enables us to probe whether LALMs succeed by genuinely processing acoustic information or by defaulting to transcript-based shortcuts.

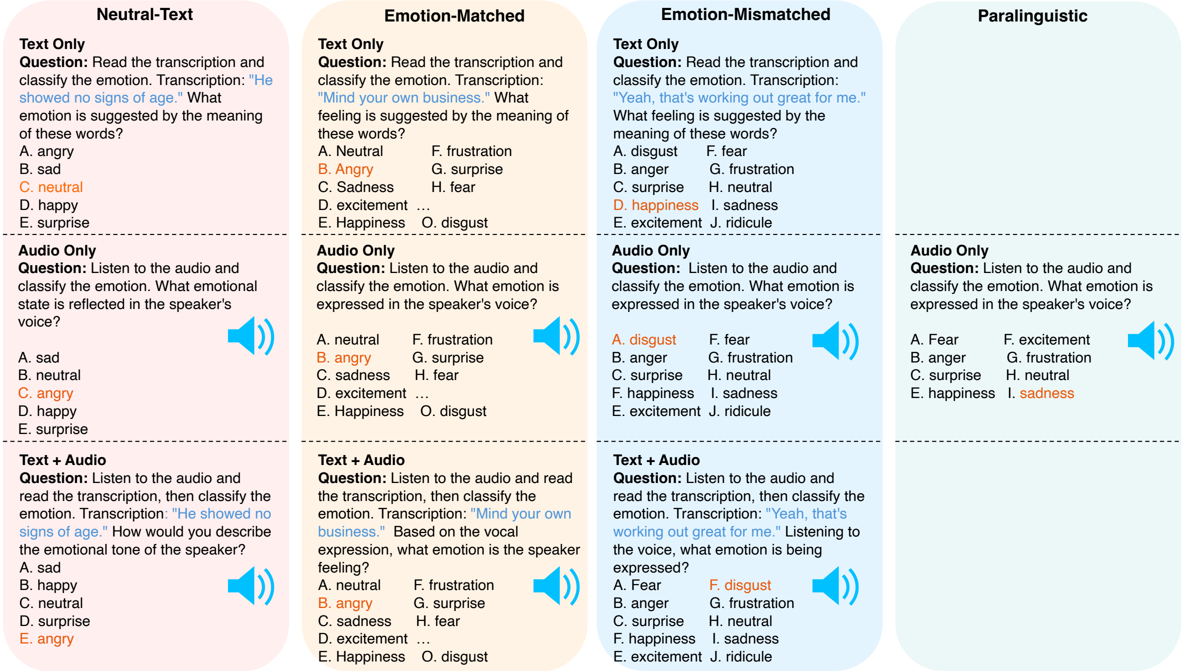
Our contributions (1) We introduce LISTEN, the first diagnostic benchmark explicitly constructed to separate lexical and acoustic effects in emotion understanding through controlled cue manipulation and multimodal evaluation. (2) We systematically evaluate six state-of-the-art open- and closed-weight LALMs across all conditions and modalities, revealing a consistent lexical dominance that limits true listening ability. (3) We analyze how cue alignment, conflict, and absence each shape model behavior, offering new insight into why current audio language models "transcribe" emotion more than they "listen" to it. LISTEN benchmark is available at: https://huggingface.co/datasets/VibeCheck1/LISTEN_full. Code is available at https://github.com/DeliJingyiC/LISTEN.

2 Related work

Audio Benchmarks Recent benchmarks have broadened LALM evaluation across diverse audio domains. MMAU (Sakshi et al., 2024) introduced 10k QA pairs over 27 skills for speech, sound, and music, while MMAU-Pro (Kumar et al., 2025) extended it to long-form, multi-source, and culturally diverse audio. MMAR (Ma et al., 2025) added 1k real-world QA triplets with hierarchical reasoning, though limited in scale. AudioBench (Wang et al., 2025a) unified 26 datasets in eight task types, and MuChoMusic (Weck et al., 2024) tested 1.1k MCQs highlighting textual bias. MMSU (Wang et al., 2025b) evaluated 5k spoken QA pairs across 47 skills, and Dynamic-SUPERB Phase-2 (yu Huang et al., 2025) covered 180 instruction-tuned speech, music, and sound tasks. AHELM (Lee et al., 2025) offered a holistic benchmark spanning perception, reasoning, emotion, bias, and multilingual safety. Although several of these benchmarks include emotion or sarcasm tasks, none directly probe how models use the speech signal itself—whether predictions arise from prosodic and paralinguistic cues or from lexical shortcuts in transcripts. Current evaluations therefore leave open a central question: to what extent do LALMs genuinely listen rather than read? Addressing this gap, LISTEN systematically manipulates lexical–acoustic alignment to disentangle transcript reliance from acoustic sensitivity, offering the first diagnostic framework for assessing whether LALMs truly process emotional information from speech.

Large Audio Language Models Recent advances in large audio language models (LALMs) such as GPT-4o (OpenAI et al., 2024), Gemini 2.5 (Comanici et al., 2025), Qwen 2.5-Omni (Xu et al., 2025a), Qwen 3-Omni (Xu et al., 2025b), and Baichuan-Omni (Li et al., 2025) have demonstrated strong capabilities in processing spoken input and generating naturalistic responses. Baichuan-Omni and the Qwen Omni series are explicitly reported to derive from text-based LLM backbones later adapted to audio through multimodal training. While this design transfers rich linguistic knowledge from text, it also risks introducing a transcript-dominant bias in which lexical information overrides acoustic evidence. Gemini 2.5, though highly multimodal, has not disclosed its training specifics, leaving similar concerns unresolved.

Speech Emotion Recognition Recent advances in deep learning have markedly improved Speech Emotion Recognition (SER). End-to-end CNN–RNN architectures can learn hierarchical acoustic representations from waveforms or spectrograms, achieving high accuracy and robustness across speakers and recording conditions (Barhoumi and BenAyed, 2024). More recent transformer and attention-based systems further enhance performance through intra-speech feature fusion, and graph-based modeling (Singh et al., 2023; Chowdhury et al., 2025). These approaches show that explicitly modeling prosodic, spectral, and temporal features yields strong emotion recognition accuracy from speech alone. Nevertheless, most SER systems remain confined to the acoustic modality and do not assess how lexical and acoustic cues jointly contribute to emotion understanding.



Text Only
Question: Read the transcription and classify the emotion. Transcription: "He showed no signs of age." What emotion is suggested by the meaning of these words?
A. angry
B. sad
C. neutral
D. happy
E. surprise

Audio Only
Question: Listen to the audio and classify the emotion. What emotional state is reflected in the speaker's voice?
A. sad
B. neutral
C. angry
D. happy
E. surprise

Text + Audio
Question: Listen to the audio and read the transcription, then classify the emotion. Transcription: "He showed no signs of age." How would you describe the emotional tone of the speaker?
A. sad
B. happy
C. neutral
D. surprise
E. angry

Text Only
Question: Read the transcription and classify the emotion. Transcription: "Mind your own business." What feeling is suggested by the meaning of these words?
A. Neutral F. frustration
B. Angry G. surprise
C. Sadness H. fear
D. excitement …
E. Happiness O. disgust

Audio Only
Question: Listen to the audio and classify the emotion. What emotion is expressed in the speaker's voice?
A. neutral F. frustration
B. angry G. surprise
C. sadness H. fear
D. excitement …
E. Happiness O. disgust

Text + Audio
Question: Listen to the audio and read the transcription, then classify the emotion. Transcription: "Mind your own business." Based on the vocal expression, what emotion is the speaker feeling?
A. neutral F. frustration
B. angry G. surprise
C. sadness H. fear
D. excitement …
E. Happiness O. disgust

Text Only
Question: Read the transcription and classify the emotion. Transcription: "Yeah, that's working out great for me." What feeling is suggested by the meaning of these words?
A. disgust F. fear
B. anger G. frustration
C. surprise H. neutral
D. happiness I. sadness
E. excitement J. ridicule

Audio Only
Question: Listen to the audio and classify the emotion. What emotion is expressed in the speaker's voice?
A. disgust F. fear
B. anger G. frustration
C. surprise H. neutral
F. happiness I. sadness
E. excitement J. ridicule

Text + Audio
Question: Listen to the audio and read the transcription, then classify the emotion. Transcription: "Yeah, that's working out great for me." Listening to the voice, what emotion is being expressed?
A. Fear F. disgust
B. anger G. frustration
C. surprise H. neutral
F. happiness I. sadness
E. excitement J. ridicule

Audio Only
Question: Listen to the audio and classify the emotion. What emotion is expressed in the speaker's voice?
A. Fear F. excitement
B. anger G. frustration
C. surprise H. neutral
E. happiness I. sadness

Figure 1. Examples from the LISTEN benchmark.

3 The LISTEN Framework

LISTEN is an evaluation framework designed to assess how large audio language models balance reliance on lexical cues versus acoustic cues in speech emotion understanding. The primary goal of LISTEN is not simply to measure classification accuracy, but to provide a controlled setting that disentangles lexical dependence from acoustic sensitivity, thereby enabling deeper insight into whether models are genuinely *listening* to speech.

The benchmark is organized into four carefully designed conditions that manipulate the alignment between lexical cues and acoustic cues: (i) **Neutral-Text**, where lexical contents are emotionally neutral but acoustic cues vary, isolating the contribution of acoustic cues; (ii) **Emotion-Matched**, where lexical and acoustic cues are aligned to reinforce the same affect; (iii) **Emotion-Mismatched**, where lexical and acoustic cues conflict, as in sarcasm; and (iv) **Paralinguistic**, where affect is conveyed without lexical content (e.g., laugh, sighs, breathing). Within the first three conditions, LISTEN further compares performance across Text-only (lexical cues only), Audio-only (acoustic and implicit lexical cues), and Text+Audio (both modalities with explicit lexical reinforcement) modalities to probe modality-specific reliance.

This design is grounded in psycholinguistic findings that lexical and acoustic channels do not contribute equally in all contexts: when the two align, both provide useful evidence, but when they conflict, as in sarcasm, acoustic cue carries the decisive signal (Bryant and Fox Tree, 2005; Attardo et al., 2003). Our goal is not to require LALMs to mimic human processing, but to test whether they can adaptively exploit the information present in speech, identifying which cues are informative and when they should be weighted more heavily. In emotion-matched situations, lexical meaning may be useful, but in mismatched or paralinguistic settings, models must rely on acoustic and nonverbal signals to reach the correct interpretation. LISTEN makes these trade-offs explicit: by manipulating lexical–acoustic alignment, it forces models to reveal whether they are genuinely leveraging speech audio or defaulting to transcript shortcuts.

3.1 Data Construction

Selected samples from the LISTEN benchmark are shown in Figure 1. Our benchmark construction follows a three-stage procedure to ensure theoretical grounding, dataset diversity, and quality control.

Stage 1: Condition Design. Building on the four experimental conditions, we formalize operational definitions and modality-specific ground-truth mapping to guide dataset selection and annotation. Neutral-Text: lexically neutral sentences with varied emotional acoustic cues; the text-only ground truth is fixed to neutral, while audio-only and text+audio use the sample's true emotion label. Emotion-Matched: lexical meaning and acoustic cues are aligned; all three modalities use the true emotion label. Emotion-Mismatched: controlled conflicts between lexical polarity and prosodic tone; text-only uses the dataset's explicit emotion label, whereas audio-only and text+audio use the dataset's implicit emotion label. Paralinguistic: no lexical content; use the true emotion label. Representative examples appear in the Appendix A.8.

Stage 2: Dataset Mapping. Each condition is instantiated using established emotional speech corpora spanning monologue and dialogue, acted and spontaneous data. Neutral-Text includes acted corpora such as CREMA-D (Cao et al., 2014), Emotion Speech Dataset (Zhou et al., 2022), TESS (Schuller et al., 2010), SAVEE (King and Narayanan, 2011), and RAVDESS (Livingstone and Brown, 2018), which are specifically designed with fixed, semantically neutral sentences (e.g., "It's eleven o'clock") spoken with varied emotional prosody, ensuring affect is conveyed solely through acoustic cues. Emotion-Matched covers both acted and spontaneous corpora, including IEMOCAP (Busso et al., 2008), CMU-MOSEI (Zadeh et al., 2018), OMGEmotionCh (Liu et al., 2021), MSP-PODCAST (Gladstone et al., 2020), and MELD (Chen et al., 2020), where lexical semantics and prosody are broadly aligned; Cue alignment for this group was verified through corpus-level metadata. Emotion-Mismatched leverages MUSTARD++ (Ray et al., 2022), which contains sarcastic and ironic speech where lexical sentiment and prosodic emotion deliberately conflict; to verify this divergence, we manually examined the explicit emotion conveyed by lexical sentiment and the implicit emotion expressed through acoustic delivery in the corpus metadata. Finally, paralinguistic samples are extracted from IEMOCAP using annotated transcripts. Additional information on each dataset can be found in Appendix A.1.

Stage 3: Question Generation. For each condition and modality (Text-only, Audio-only, Text+Audio), standardized multiple-choice emotion recognition prompts were generated using GPT-5 (OpenAI, 2025). Parallel templates targeted the same judgment, identifying the speaker's emotion, while differing in focus on lexical meaning, vocal prosody, or both. To minimize inference-time bias, one paraphrased question form was randomly selected per item from a small pool of semantically equivalent variants, and answer options were shuffled to prevent positional bias. All prompt templates were manually reviewed to ensure semantic equivalence and modality relevance. Representative examples are provided in Appendix A.2.

3.2 LISTEN Statistics

Table 1 summarizes the core statistics of the LISTEN benchmark, which consists of 7,939 evaluation questions spanning four experimental conditions and three modality partitions. Among the four conditions, *Neutral-Text* constitutes the largest portion with 3,428 questions, followed by *Emotion-Matched* (3,155), *Paralinguistic* (975), and *Emotion-Mismatched* (381). On average, each question contains 14.7 words, while answer options average 8.3 words. The associated audio clips are short and focused, averaging 3.5 seconds, ensuring that emotional cues remain perceptually salient without introducing long-context confounds. Detailed distributions for each dataset and condition are provided in the Appendix A.3.

Table 1. Key statistics of the LISTEN benchmark.

Statistic	**Value**
Total questions	7,939
Task count	4
Modality count	3
Neutral-Text	3,428
Emotion-Matched	3,155
Emotion-Mismatched	381
Paralinguistic	975
Average question length	14.7 words
Average option length	8.3 words
Average audio length	3.5 seconds

4 Experiments

Models We evaluate state-of-the-art large audio language models including closed-weight models, Gemini 2.5 Flash and Gemini 2.5 Pro (Comanici et al., 2025); open-weight models: Qwen2.5-Omni-7B (Xu et al., 2025a), Qwen3-Omni-30B (Xu et al., 2025b), Baichuan-Omni-1.5 (Li et al., 2025), and Qwen3-Instruct (Xu et al., 2025b). The hyperparameters and configurations used during the evaluation process are consistent with their official settings. Details appear in Appendix A.4.

Evaluation Protocol For each sample, we present the model with a multiple-choice question containing 5-10 emotion options (happiness, sadness, anger, fear, surprise, disgust, neutral, frustration, excitement, ridicule). To prevent position bias, we randomize the order of choices for each query. Models receive only the audio, text, or both depending on the condition, with sample identifiers anonymized to prevent information leakage.

We employ zero-shot evaluation with carefully designed prompts that instruct models to classify emotions based solely on the provided input. For text-only conditions, models receive transcriptions without timing or acoustic annotations. For audio-only conditions, models process raw audio without text transcripts. For multimodal conditions, both audio and text transcripts are provided simultaneously. Examples are provided in Appendix A.8.

4.1 Metrics

We evaluate model performance using overall **accuracy**, the proportion of correctly classified samples across emotion categories. Since each sample is assigned a single ground-truth label, accuracy reflects the correctness and equals micro-F1 in this single-label multi-class setting. To interpret model performance relative to chance, we report three reference baselines for each experiment:

- **Uniform Guess:** assumes a random classifier that predicts each emotion class with equal probability. This baseline represents the accuracy expected from purely random guessing with no prior information about the dataset.

- **Majority Guess:** always predicts the most frequent emotion in the dataset. This provides an upper bound for trivial label-frequency heuristics and reflects dataset imbalance.

- **Prediction-Marginal Distribution Baseline:** estimates the expected accuracy of a random classifier that samples predictions according to the model's own empirical prediction distribution rather than uniformly. This captures how much of a model's performance can be attributed to its output bias rather than meaningful input sensitivity.

Formally, let p_i denote the probability the model predicts class i, and q_i the empirical frequency of class i in the ground-truth labels. The expected prediction-marginal accuracy is computed as:

$$\mathbb{E}[\text{Acc}] = \sum_i p_i q_i. \quad (1)$$

This expectation accounts for both dataset imbalance (q_i) and model prediction bias (p_i), yielding a more informative lower bound than uniform guessing. When $p_i = q_i$, the expected value corresponds to the Bayes-optimal random baseline given the dataset label prior. We report all accuracies in Table 5, with each model's prediction-marginal baseline shown in parentheses for direct comparison. The gaps between the accuracy reported and the prediction marginal distribution baseline reveal how much performance comes from the interpretation of input cues rather than from model's own prediction bias. Larger gaps indicate stronger use of real lexical or acoustic information.

In addition to accuracy, we report Unweighted Average Recall (UAR) and Macro-F1 to account for class imbalance. A compact summary is provided in the appendix (see 4), with full per-model and per-condition tables.

5 Results and Analysis

Overall Performance Table 5 and Figure 2 summarizes results across all experimental conditions. Overall, models still rely more on text than on acoustic cues, but their behavior shifts depending on how speech and lexical cues align.

Neutral-Text condition the transcripts are deliberately emotionless while the audio expresses a range of emotions, allowing us to isolate the role of acoustic cues. In the text-only setting, where all transcripts are neutral and the ground-truth label is neutral for every sample, LALMs reach near-perfect accuracy (e.g., 96.6% for Gemini2.5-Pro, 85.4% for the Qwen Omnis). In contrast, in audio-only setting, models must infer emotion directly from acoustic cues. It shows much lower scores (25–35%), revealing the difficulty of recognizing emotional tone from acoustic cues. Qwen2.5-Omni-7B and Gemini 2.5-Pro has 34.0% and 34.9% accuracy relatively. Other models' accuracy are all below 30.0% When text and audio are combined, the closed-weight Gemini2.5-Pro shows a modest gain (41.7%), suggesting partial use of acoustic cues even when lexical content is neutral, whereas open-weight models (e.g., Qwen2.5-Omni-7B, Baichuan-Omni-1.5) and Gemini2.5-Flash often perform worse than their audio-only baselines.

Table 2. Accuracy (with prediction marginal distribution baseline in parenthesis—see subsection 4.1) are reported. **Bold** values highlight the highest value and underlined values highlight the second-highest value in each experiment. For each condition, the Average column represents the mean of the audio and text+audio settings. The Overall Average is computed as the mean accuracy across all audio and text+audio results from the four experimental conditions (seven modalities in total).

	Neutral-Text				Emotion-Matched				Emotion-Mismatched				Paralinguistic	Overall
Model	Text	Audio	Text+Audio	Average	Text	Audio	Text+Audio	Average	Text	Audio	Text+Audio	Average	Audio	Average
Uniform Guess	12.5	12.5	12.5	12.5	6.7	6.7	6.7	6.7	10.0	10.0	10.0	10.0	12.5	10.1
Majority Guess	100	16.9	16.9	16.9	26.5	26.5	26.5	26.5	39.0	39.0	39.0	39.0	32.6	28.2
							Open-Weight Model							
Qwen3-instruct	66.6 (65.0)	–	–	–	33.5 (12.4)	–	–	–	**38.0** (29.8)	–	–	–	–	–
Qwen2.5-Omni-7B	85.4 (84.1)	34.0 (10.9)	19.8 (11.8)	26.9	36.4 (15.7)	36.6 (14.1)	38.6 (15.7)	37.6	34.0 (13.5)	**38.5** (29.5)	39.1 (23.6)	38.8	**22.7** (11.4)	32.8
Qwen3-Omni-30B	85.4 (84.0)	29.3 (11.3)	25.3 (11.8)	27.3	38.7 (11.7)	**42.4** (15.8)	**43.1** (15.5)	**42.8**	34.6 (12.2)	37.4 (26.7)	39.1 (25.2)	38.3	21.0 (12.6)	33.9
Baichuan-Omni-1.5	81.2 (79.6)	16.5 (11.5)	15.2 (12.1)	15.9	31.0 (15.9)	36.0 (15.6)	36.0 (17.8)	36.0	31.0 (27.5)	36.0 (31.0)	36.0 (32.0)	36.0	**22.7** (11.5)	28.3
							Closed-Weight Model							
Gemini2.5-Flash	82.5 (81.5)	25.6 (10.7)	24.6 (11.0)	25.1	36.6 (14.3)	30.7 (12.3)	38.9 (14.1)	34.8	33.1 (10.1)	35.8 (15.5)	38.0 (18.5)	36.9	18.0 (12.0)	30.2
Gemini2.5-Pro	**96.6** (96.6)	**34.9** (12.8)	**41.7** (12.9)	**38.3**	**38.8** (15.7)	37.6 (16.9)	40.2 (14.3)	38.9	31.8 (10.0)	36.9 (22.5)	**42.6** (23.3)	**39.8**	15.7 (9.3)	**35.7**

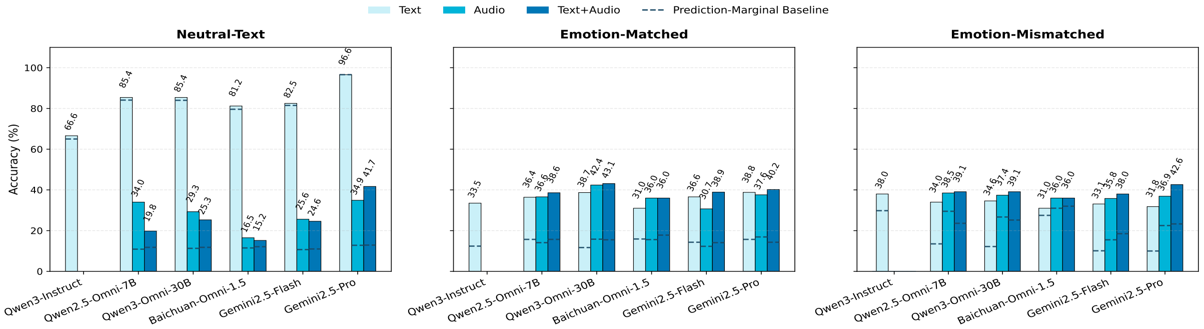


Figure 2. Model accuracy across three LISTEN conditions (Neutral-Text, Emotion-Matched, Emotion-Mismatched) under text-only, audio-only, and text+audio modalities. Dashed lines indicate prediction-marginal baselines.

Qwen2.5-Omni-7B has a large accuracy drop from audio only to text+audio (34.0% → 19.8%)

Further evidence of this pattern is shown in the confusion matrices in Figure 3, which visualizes Gemini 2.5-Pro's emotion-recognition behavior across the three LISTEN conditions—(1) Neutral-Text, (2) Emotion-Matched, and (3) Emotion-Mismatched—each tested with text-only, audio-only, and audio + text inputs. The top row three heat maps corresponds to the Neutral-Text condition. In the text-only setting, the confusion matrix shows a single dominant bar in the neutral column, indicating that the model predicts "neutral" for nearly all samples. In both the audio-only and audio + text settings, a faint diagonal emerges, showing that the model can identify a subset of emotional classes from acoustic cues. However, the same vertical concentration in the neutral column persists, showing that the neutral lexical content biases the model's interpretation even when the speech conveys clear emotional tone. Overall, this pattern indicates that LALMs over-emphasize lexical cues, leading to interference rather than effective identify emotion based on acoustic cues. More details are shown in Figure 8.

Emotion-Matched condition. In this setting, the speech and transcripts express the same emotion, allowing us to evaluate how LALMs process consistent lexical–acoustic alignment. Performance across models is relatively balanced between text and audio inputs, with modest multimodal gains suggesting limited but consistent cue integration. Among open-weight systems, Qwen3-Omni-30B achieves the highest scores: 38. 7% (text only), 42. 4% (audio only) and 43. 1% (text + audio), showing that it effectively leverages both lexical and acoustic cues when they match. Qwen2.5-Omni-7B follows a similar pattern (36.4–38.6%), while Baichuan-Omni-1.5 performs consistently lower (31–36%), suggesting a more limited capacity for recognizing emotions from speech.

For closed-weight models, Gemini 2.5-Pro performs strongly and consistently across modalities: 38.8% (text-only), 37. 6% (audio-only) and 40.2% (text + audio) demonstrating stable use of lexical and acoustic cues. In contrast, Gemini 2.5-Flash performs worse when using only audio: its accuracy drops from 36.6% with text to 30.7% with audio, showing that it struggles to capture emotional tone even when acoustic cues and lexical cues have matched emotion in the audio. When both inputs are combined, its score rises again to

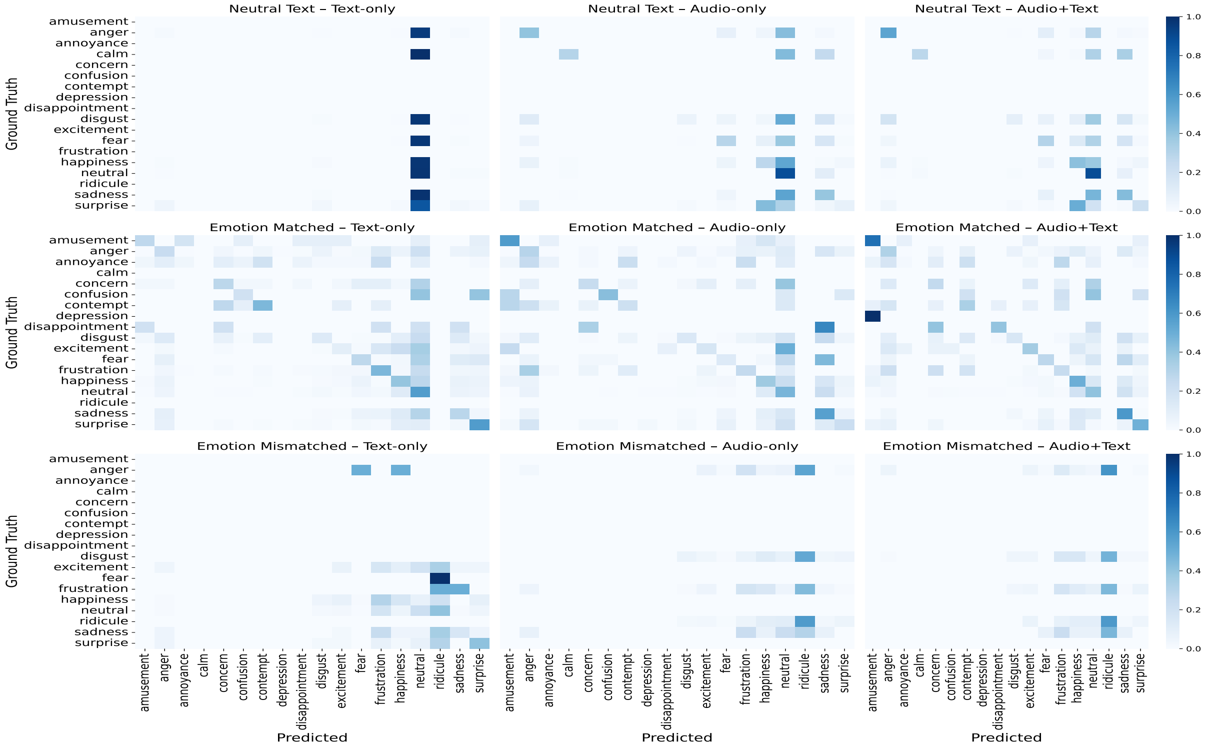


Figure 3. Confusion matrices showing Gemini 2.5 Pro's emotion recognition performance across three experimental conditions in the LISTEN benchmark. Row-normalized matrices display prediction distributions for each true emotion class across: (1) Neutral Text, (2) Emotion Matched, and (3) Emotion Mismatched conditions, each tested with text-only, audio-only, and audio+text modalities.

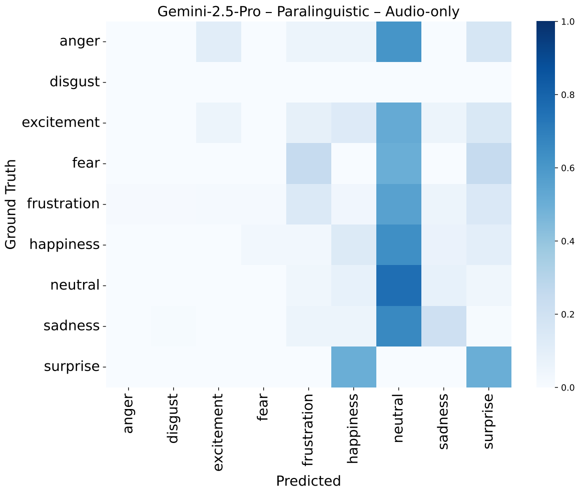


Figure 4. Confusion matrices showing Gemini 2.5 Pro's emotion recognition performance in paralinguistic condition

38.9%, suggesting that adding text helps the model recover accuracy by providing a more stable signal.

These trends are illustrated in Figure 3, where the three matrices in the second row (Emotion-Matched condition) for Gemini 2.5 Pro display a faint diagonal.In the text-only and audio-only settings, a visible neutral column remains, indicating some bias toward neutral predictions. In text+audio setting, the diagonal becomes slightly clearer and the neutral column weakens, indicating that emphasizing lexical content enhances discriminability across emotion categories when textual and acoustic signals are aligned. Overall, although most LALMs can recognize emotion better when lexical and acoustic cues point to the same label, compared to neutral-text, lexical emotion recognition (text-only) is still weak, and combining lexical and acoustic acoustic emotion recognition (audio-only, text+audio) yields little improvement. Some architectures (like Gemini 2.5-Flash) still struggle to exploit acoustic information effectively. Confusion matrices for all models are shown in Figure 9.

Emotion-Mismatched condition. In this condition, the speech and text express opposing emotions, testing how LALMs weight lexical vs. acoustic cues when they conflict. The results suggest that LALMs can detect sarcasm when lexical and acoustic cues disagree but struggle to identify the specific sarcastic emotion underlying that mismatch.

Among open-weight models, both Qwen2.5-Omni-7B (34.0% → 38.5% → 39.1%) and Qwen3-Omni-30B (34.6% → 37.4% → 39.1%) gain modest improvements from audio. Baichuan-Omni-

1.5 (31.0% → 36.0% → 36.0%) has lower accuracy across all three settings, and the difference between these accuracy and the prediction marginal distribution baseline is small, suggesting limited discriminative use of input signals. For closed-weight models, both Gemini2.5-Flash (33.1% → 35.8% → 38.0%) and Gemini2.5-Pro (31.8% → 36.9% → 42.6%) show clearer acoustic utilization. Gemini2.5-Pro achieves 42.6% in the text+audio setting versus a 23.3% prediction-marginal baseline, demonstrating better sensitivity to acoustic cues under cue conflict.

However, the overall gaps between reported accuracies and prediction marginal distribution baseline are smaller here than in the Neutral-Text or Emotion-Matched conditions, implying that much of the observed performance may come from prediction biases toward a few dominant emotional categories rather than fine-grained emotion recognition in lexical and acoustic cues. The heat maps in Figure 3 support this interpretation: in the Emotion-Mismatched condition (bottom row), the text-only setting shows weak diagonals concentrated around a few emotions, mainly happiness, neutral, surprise, and sadness. However, both the audio-only and text+audio settings show strong vertical bars on ridicule and, to a lesser extent, frustration, indicating that models classify a large portion of conflicting samples into these categories. This pattern suggests that LALMs recognize the presence of emotional conflict in lexical and acoustic cues but resolve it by collapsing diverse sarcastic emotions (anger, disgust, frustration, ridicule, sadness) into only two classes: ridicule and frustration, revealing a key limitation in current LALM understanding of complex sarcastic emotions. We observe similar performance in Qwen2.5-omni, Qwen3-omni, Baichuan-omni, and Gemini2.5-Flash. More details are shown in Figure 10.

Paralinguistic condition. This condition isolates nonlexical affective cues, such as laughter, sighs, gasps, or other vocalizations, by removing linguistic content entirely. Performance across models are just above uniform guess baseline and prediction marginal distribution baseline, showing that current LALMs have limited ability to interpret emotion from nonverbal sounds alone. Among open-weight systems, Qwen2.5-Omni-7B and Baichuan-Omni-1.5 reach the highest accuracy (22.7%), followed by Qwen3-Omni-30B (21.0%), while the closed-weight Gemini2.5-Flash (18.0%) and Gemini2.5-Pro (15.7%) perform slightly lower. The confusion matrix for Gemini2.5-Pro in Figure 3 illustrates this limitation. A faint diagonal suggests partial recognition of emotions like surprise and sadness, while the dominant neutral column reveals a strong bias toward predicting “neutral” for nearly all categories except "surprise". This bias shows limited ability to distinguish nonverbal emotions and reliance on lexical cues for emotion identification. See Appendix A.6, for detailed model results.

6 Discussion and Conclusion

Where Do Models Succeed and Fail? Lexical Dominance and the Limits of Listening. LISTEN reveals a mixed and fragile affective skill profile in current LALMs. In the Neutral-Text, Audio-only, and Text+Audio conditions, models remain strongly biased toward predicting “neutral,” reflecting an overreliance on lexical cues and a tendency to default to neutral interpretations regardless of acoustic cues. In the Paralinguistic condition, where no lexical content is available at all, models still default to “neutral,” not because of lexical interference, but because of the absence of lexical grounding. This pattern suggests that current LALMs depend on textual information both as an interpretive anchor and as a confidence cue: when lexical guidance is strong, they ignore acoustic variation, and when it is missing, they revert to neutral as the safest default.

When lexical and acoustic cues align, as in the Emotion-Matched condition, models can recognize a wider range of emotions, but overall accuracy remains low and the neutral bias persists. This suggests that their emotion recognition benefits only marginally from cue consistency and that effective multimodal integration is still limited in LALMs.

When lexical and acoustic cues conflict, as in the Emotion-Mismatched condition, LALMs can detect that an emotional discrepancy exists but fail to interpret it precisely. They can identify such conflicts as sarcasm, collapsing diverse sarcastic emotional states (e.g., anger, disgust, or ridicule) into a narrow set of categories. This indicates that while models sense the presence of incongruity between modalities, they still lack the ability to reason about the nuanced emotional or pragmatic meaning behind that conflict.

Are Fine-Grained Emotion Labels a Source of Noise?

Are Fine-Grained Emotion Labels a Source of Noise? A potential concern in comparing fine- and coarse-grained evaluations is whether increased label granularity introduces annotation ambiguity that inflates error rates. Our results suggest this is not the case. Across most conditions, performance differences between fine- and coarse-grained evaluation are modest, with only minor changes in the Neutral-Text and Emotion-Matched settings. Analysis of close-category confusions (e.g., frustration vs. anger) shows that such errors account for only a small fraction of total mistakes, indicating that fine-grained labels are not a major source of noise. For example, in the Emotion-Matched, Audio-only condition for Gemini 2.5-Pro, neighboring confusions (e.g., anger–frustration, excitement–happiness) constitute a minority of errors.

The primary exception is the Emotion-Mismatched condition, where coarse-grained accuracy increases substantially. This increase is condition-specific rather than global and arises from the interaction between label aggregation and model behavior. Sarcastic utterances span multiple fine-grained emotion labels that are merged into a single coarse category, and when faced with conflicting lexical and acoustic cues, models tend to default to a small subset of emotions, most notably frustration and ridicule. Once these labels are collapsed, many such predictions become correct by construction (see Appendix subsection A.5). Together, these findings indicate that fine-grained emotion labels do not inflate error rates; instead, they reveal systematic model limitations that are obscured under coarse-grained evaluation.

Lexical vs. Acoustic Cue Following Under Conflict. To quantify how models resolve conflicts between lexical and acoustic signals, we analyzed Gemini 2.5 Pro's predictions in the Emotion-Mismatched condition and measured how often outputs align with the lexical versus the acoustic cue. In the Audio-only condition, 96 out of 381 predictions align with the acoustic emotion, compared to 20 out of 381 that align with the lexical cue. A similar pattern is observed in the Text+Audio condition, where 93/381 predictions align with the acoustic cue and 24/381 with the lexical cue. These results indicate that current LALMs can identify cross-modal emotional conflict and, in a subset of cases, preferentially rely on acoustic information. However, this behavior remains limited and inconsistent: rather than resolving the conflict by accurately inferring the intended emotional state, models tend to collapse diverse conflicting signals into a narrow set of categories, predominantly ridicule and frustration. Overall, this pattern suggests emerging sensitivity to acoustic cues without robust, fine-grained conflict resolution.

Why Not Use SER Models as Primary Baselines. We agree that standard speech emotion recognition (SER) systems can be informative, and we conducted preliminary experiments with representative SER baselines. The strong in-domain performance of task-specific SER models demonstrates that the acoustic signal contains sufficient information for accurate emotion recognition under matched conditions, highlighting the gap between current LALMs and specialized speech systems.

However, SER models are constrained by fixed label spaces and dataset-specific training, making them incompatible with LISTEN's multi-dataset design and controlled manipulation of lexical–acoustic alignment. For example, a SpeechBrain wav2vec2 SER model trained on IEMOCAP achieves 77% accuracy on the IEMOCAP subset of LISTEN but only 22% on RAVDESS, revealing substantial domain sensitivity. Thus, while SER models serve as useful task-specific upper bounds, they are not suitable as primary baselines for LISTEN's controlled evaluation setting.

Implications and Future Directions. In general, LISTEN highlights a central finding behind our question: Audio LLMs often transcribe more than they truly listen. While they can detect emotional variation in speech, their interpretations remain shallow and heavily guided by lexical information. Even when provided with rich acoustic cues, models tend to default to "neutral" predictions, revealing limited prosodic sensitivity and weak integration between text and audio streams. Future progress in "listening" models may benefit from the advances achieved in corporate and state-of-the-art speech emotion recognition (SER) systems, which attain strong accuracy by explicitly modeling prosodic, spectral, and temporal features of speech. Techniques on emotional speech could be adapted to improve prosodic grounding and acoustic sensitivity in LALMs. Building on these insights, next-generation audio language models should not only perceive acoustic variation but also infer its emotional and communicative intent.

Limitations

While LISTEN is designed to provide a controlled and interpretable diagnostic of lexical versus acoustic cue reliance in LALMs, several scope limitations merit discussion. First, each utterance is evaluated in isolation rather than within a multi-turn conversational context. This design choice enables precise manipulation of lexical–acoustic alignment and isolates local emotional expression, but it omits dialogue history and sequential emotion flow. Incorporating conversational context may alter how models balance acoustic and lexical cues, an open question that future extensions of LISTEN are well-positioned to explore.

Second, the benchmark is currently restricted to English-language datasets. This limits the assessment of cross-linguistic generalizability, particularly given well-documented variation in how emotional prosody and acoustic cues are realized across languages. Expanding LISTEN to multilingual settings would enable systematic study of language-dependent acoustic–lexical interactions and improve applicability to diverse real-world scenarios.

Ethical Considerations

All datasets used in this work are publicly available and were originally collected under ethical or open-use guidelines. LISTEN is designed solely for research evaluation and does not involve new human data collection. We acknowledge that emotion recognition technologies carry potential privacy and misuse risks, and encourage future work to prioritize transparency, fairness, and responsible deployment.

We are thankful for the generous support of computational resources provided by the Ohio Supercomputer Center.

References

Salvatore Attardo, Jodi Eisterhold, Jennifery Hay, and Isabella Poggi. 2003. Multimodal markers of irony and sarcasm. *Humor: International Journal of Humor Research*, 16(2).

Rainer Banse and Klaus R Scherer. 1996. Acoustic profiles in vocal emotion expression. *Journal of personality and social psychology*, 70(3):614.

Chawki Barhoumi and Yassine BenAyed. 2024. Real-time speech emotion recognition using deep learning and data augmentation. *Artificial Intelligence Review*, 58(2):49.

Gregory A Bryant and Jean E Fox Tree. 2005. Is there an ironic tone of voice? *Language and speech*, 48(3):257–277.

Carlos Busso, Murtaza Bulut, Chin-Hui Lee, and Shrikanth Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4):335–359.

Huan Cao, Carlos Busso, and Chin-Hui Lee. 2014. Crema-d: A corpus of realistic emotional speech. *IEEE Transactions on Affective Computing*, 5(4):441–454.

Ming Chen and 1 others. 2020. Mels: A multimodal emotional speech dataset. *IEEE Transactions on Affective Computing*.

Jaher Hassan Chowdhury, Sheela Ramanna, and Ketan Kotecha. 2025. Speech emotion recognition with light weight deep neural ensemble model using hand crafted features. *Scientific Reports*, 15(1):11824.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3290 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

Brittany Gladstone and 1 others. 2020. Msu-podcast: A large-scale naturalistic podcast dataset for emotion recognition. *IEEE Transactions on Affective Computing*.

Seung-Goo King and Shrikanth Narayanan. 2011. The savee database of emotional speech. *Proceedings of Interspeech*.

Sonal Kumar, Šimon Sedláček, Vaibhavi Lokegaonkar, Fernando López, Wenyi Yu, Nishit Anand, Hyeonggon Ryu, Lichang Chen, Maxim Plička, Miroslav Hlaváček, William Fineas Ellingwood, Sathvik Udupa, Siyuan Hou, Allison Ferner, Sara Barahona, Cecilia Bolaños, Satish Rahi, Laura Herrera-Alarcón, Satvik Dixit, and 15 others. 2025. Mmau-pro: A challenging and comprehensive benchmark for holistic evaluation of audio general intelligence. *Preprint*, arXiv:2508.13992.

Tony Lee, Haoqin Tu, Chi Heem Wong, Zijun Wang, Siwei Yang, Yifan Mai, Yuyin Zhou, Cihang Xie, and Percy Liang. 2025. Ahelm: A holistic evaluation of audio-language models. *arXiv preprint arXiv:2508.21376*.

Yadong Li, Jun Liu, Tao Zhang, Tao Zhang, Song Chen, Tianpeng Li, Zehuan Li, Lijun Liu, Lingfeng Ming, Guosheng Dong, Da Pan, Chong Li, Yuanbo Fang, Dongdong Kuang, Mingrui Wang, Chenglin Zhu, Youwei Zhang, Hongyu Guo, Fengyu Zhang, and 74 others. 2025. Baichuan-omni-1.5 technical report. *Preprint*, arXiv:2501.15368.

Guolei Liu and 1 others. 2021. Omgemotionchallenge: A challenge dataset for emotion recognition in videos. *IEEE Transactions on Affective Computing*.

Stephen Livingstone and Mark Brown. 2018. The ravdess emotional speech and song dataset. In *Proceedings of the 9th International Conference on Affective Computing and Intelligent Interaction*.

Ziyang Ma, Yinghao Ma, Yanqiao Zhu, Chen Yang, Yi-Wen Chao, Ruiyang Xu, Wenxi Chen, Yuanzhe Chen, Zhuo Chen, Jian Cong, Kai Li, Keliang Li, Siyou Li, Xinfeng Li, Xiquan Li, Zheng Lian, Yuzhe Liang, Minghao Liu, Zhikang Niu, and 15 others. 2025. Mmar: A challenging benchmark for deep reasoning in speech, audio, music, and their mix. *Preprint*, arXiv:2505.13032.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

OpenAI. 2025. Gpt-5 system card.

Amit Ray and 1 others. 2022. Mustard++: Multimodal sarcasm detection with extended emotion labels. In *Proceedings of the International Conference on Multimodal Interaction*.

S Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. 2024. Mmau: A massive multi-task audio understanding and reasoning benchmark. *Preprint*, arXiv:2410.19168.

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

Björn Schuller, Stephan Steidl, Anton Batliner, Dietrich Seppi, Klaus Kroschel, and Gerhard Rigoll. 2010. The tess corpus of emotional speech. *Proceedings of LREC*, 1:3123–3127.

Jagjeet Singh, Lakshmi Babu Saheer, and Oliver Faust. 2023. Speech emotion recognition using attention model. *International Journal of Environmental Research and Public Health*, 20(6):5140.

Bin Wang, Xunlong Zou, Geyu Lin, Shuo Sun, Zhuohan Liu, Wenyu Zhang, Zhengyuan Liu, AiTi Aw, and Nancy F. Chen. 2025a. AudioBench: A universal benchmark for audio large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4297–4316, Albuquerque, New Mexico. Association for Computational Linguistics.

Dingdong Wang, Jincenzi Wu, Junan Li, Dongchao Yang, Xueyuan Chen, Tianhua Zhang, and Helen Meng. 2025b. Mmsu: A massive multi-task spoken language understanding and reasoning benchmark. *arXiv preprint arXiv:2506.04779*.

Benno Weck, Ilaria Manco, Emmanouil Benetos, Elio Quinton, George Fazekas, and Dmitry Bogdanov. 2024. Muchomusic: Evaluating music understanding in multimodal audio-language models. *Preprint*, arXiv:2408.01337.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. 2025a. Qwen2.5-omni technical report. *Preprint*, arXiv:2503.20215.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, and 19 others. 2025b. Qwen3-omni technical report. *Preprint*, arXiv:2509.17765.

Chien yu Huang, Wei-Chih Chen, Shu wen Yang, Andy T. Liu, Chen-An Li, Yu-Xiang Lin, Wei-Cheng Tseng, Anuj Diwan, Yi-Jen Shih, Jiatong Shi, William Chen, Chih-Kai Yang, Wenze Ren, Xuanjun Chen, Chi-Yuan Hsiao, Puyuan Peng, Shih-Heng Wang, Chun-Yi Kuan, Ke-Han Lu, and 61 others. 2025. Dynamic-superb phase-2: A collaboratively expanding benchmark for measuring the capabilities of spoken language models with 180 tasks. *Preprint*, arXiv:2411.05361.

Amir H. Zadeh, Ming Chen, Soujanya Poria, and Louis-Philippe Morency. 2018. Cmu-mosei: A multimodal sequence-to-sequence dataset for emotion recognition. In *Proceedings of the 27th International Conference on Computational Linguistics*.

Kun Zhou, Berrak Sisman, Rui Liu, and Haizhou Li. 2022. Emotional voice conversion: Theory, databases and esd. *Speech Communication*, 137:1–18.

A Appendix

A.1 Source Datasets Used During Data Construction

Each experimental condition in our benchmark is instantiated using established emotional speech corpora spanning acted and spontaneous data, monologue and dialogue, and both congruent and incongruent emotion–text alignments.

CREMA-D (Cao et al., 2014): The Crowd-Sourced Emotional Multimodal Actors Dataset consists of 7,442 clips from 91 actors (48 male, 43 female) portraying six emotions—anger, disgust, fear, happy, neutral, and sad—across 12 fixed sentences. Acted speech. The CREMA-D dataset is licensed under the Open Database License (ODbL v1.0) (Open Data Commons). This license permits sharing, use, and adaptation, but requires attribution and that derivative databases remain under the same license.

Emotion Speech Dataset (Zhou et al., 2022): This dataset contains 350 parallel utterances spoken by 10 native Mandarin speakers, and 10 English speakers with 5 emotional states (neutral, happy, angry, sad and surprise) We only use English data. It provides consistent lexical content for analyzing prosodic variations in acted emotional speech. The RAVDESS is released under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, CC BY-NC-SA 4.0

TESS (Schuller et al., 2010): The Toronto Emotional Speech Set contains 2,800 recordings of two female actors simulating seven emotions—anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral—based on the same lexical template "Say the word ___." It is designed for perceptual studies of emotional prosody. Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND 4.0)

SAVEE (King and Narayanan, 2011): The Surrey Audio-Visual Expressed Emotion dataset features 480 utterances from four male native English speakers—DC, JE, JK, and KL—who were postgraduate students at the University of Surrey, aged between 27 and 31. The dataset encompasses seven emotional classes: anger, disgust, fear, happiness, sadness, surprise, and neutrality, as commonly defined in psychological studies. Each speaker articulated 15 sentences per emotion, selected from the TIMIT corpus. This set included three sentences shared across all emotions, two that were emotion-specific, and ten general utterances tailored to each emotion, arranged in alphabetical order. License: Data files @ Original Authors

RAVDESS (Livingstone and Brown, 2018): The Ryerson Audio-Visual Database of Emotional Speech and Song includes 7,356 audio and video clips from 24 professional actors expressing emotions through both speech and song. It encompasses calm, happy, sad, angry, fearful, surprise, and disgust, supporting multimodal emotion recognition. Licensed under CC BY-NA-SC 4.0.

IEMOCAP (Busso et al., 2008): The Interactive Emotional Dyadic Motion Capture Database comprises approximately 12 hours of audiovisual recordings from ten actors performing scripted and improvised dialogues. It is annotated for categorical emotions (anger, happiness, sadness, neutral, excitement) and dimensional affect ratings (valence, arousal, dominance), enabling analysis of spontaneous emotional dynamics. License: https://sail.usc.edu/iemocap/Data_Release_Form_IEMOCAP.pdf

CMU-MOSEI (Zadeh et al., 2018): The Multimodal Opinion Sentiment and Emotion Intensity dataset contains 23,454 sentence-level video segments from online monologues annotated for both sentiment and emotion. It provides large-scale multimodal coverage of spontaneous speech in natural contexts. License: CC BY-SA 4.0 All data copyright: Carnegie Mellon University & authors

OMG-Emotion Challenge Dataset (Liu et al., 2021): This dataset includes monologue videos annotated for continuous emotion dimensions (valence and arousal), focusing on gradual emotion evolution within a single speaker. It emphasizes contextual and temporal modeling of affective expression. License: Apache License 2.0

MSP-Podcast (Gladstone et al., 2020): A large-scale corpus of natural English speech extracted from public podcasts, containing over 100,000 segments annotated for categorical emotions and continuous dimensions. It captures rich acoustic variability and spontaneous emotional speech. This dataset has Common Licenses that permit the distribution of the corpus.

MELD (Chen et al., 2020): The Multimodal EmotionLines Dataset extends the EmotionLines corpus with audio and visual modalities from the TV series Friends. It includes 13,000 utterances from 1,433 dialogues annotated with seven emotions: anger, disgust, sadness, joy, neutral, surprise, and fear. GNU General Public License v3.0

MUSTARD++ (Ray et al., 2022): The Multimodal Sarcasm Detection dataset extends the original MUSTARD with additional sarcastic and non-sarcastic clips from TV shows. It contains audio, visual, and textual modalities annotated for sarcasm, where lexical and prosodic cues intentionally conflict, making it ideal for evaluating multimodal incongruence. The data provides both explicit emotion annotation, aligned with lexical content, and implicit emotion annotation aligned with speech. License. CC0: Public Domain.

A.2 Question Prompts

To ensure consistent evaluation across modalities, we design parallel sets of natural-language prompts for Text-only, Audio-only, and Text+Audio conditions. Below are representative examples for each modality.

Text-only

- Based on the content of this text, what emotion would the person likely be feeling?
- What emotion is conveyed by the words in this statement?
- Reading this text, what emotional state does the speaker appear to be in?
- From the semantic content alone, what emotion is being expressed?
- What feeling is suggested by the meaning of these words?
- Based solely on the text content, what emotion would you identify?
- What emotional tone is conveyed by the literal meaning of this statement?

Audio-only

- What emotion is expressed in the speaker's voice?
- What emotion does the speaker convey through their tone?
- Based on the vocal expression, what emotion is the speaker feeling?
- What emotional state is reflected in the speaker's voice?
- How would you describe the emotional tone of the speaker?
- What emotion is communicated through the speaker's vocal prosody?
- Listening to the voice, what emotion is being expressed?

Text + Audio

- Considering both the words and how they are spoken, what is the speaker's true emotional state?
- What emotion is conveyed when you combine the text content with the vocal expression?
- Based on both the semantic meaning and prosodic cues, what emotion is the speaker feeling?
- How do the words and vocal tone together reveal the speaker's emotional state?
- What emotion emerges when you integrate the textual and acoustic information?
- Taking into account both content and delivery, what emotion is being expressed?
- What is the complete emotional picture when combining words and voice?

A.3 LISTEN Emotion Distribution across experimental conditions

We visualize the distribution of ground-truth emotion labels for each experimental condition.

Neutral-Text, Emotion-Matched, and Paralinguistic Figure 5 aggregates three subplots, each showing per-dataset label counts using that dataset's own taxonomy (after consistent normalization). X-axes list only the emotions present in each dataset; the shared Y-axis facilitates cross-condition comparison of prevalence.

Emotion-Mismatched (explicit) and Emotion-Mismatched (implicit) Figure 6 isolates the two mismatched conditions, highlighting how label prevalence differs between explicit and implicit mismatches.

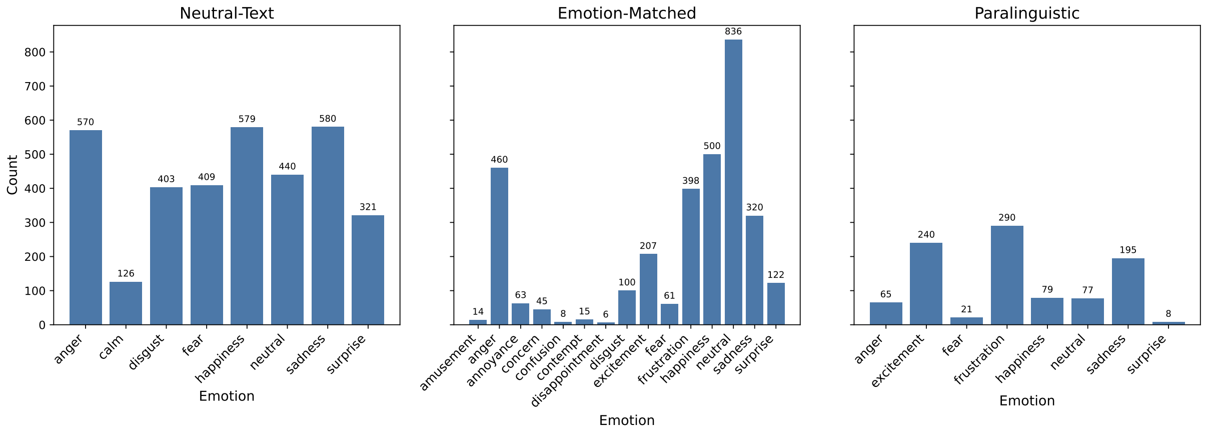


Figure 5. Ground-truth label distributions for Neutral-Text, Emotion-Matched, and Paralinguistic conditions. Each subplot shows counts for the labels present in that dataset.

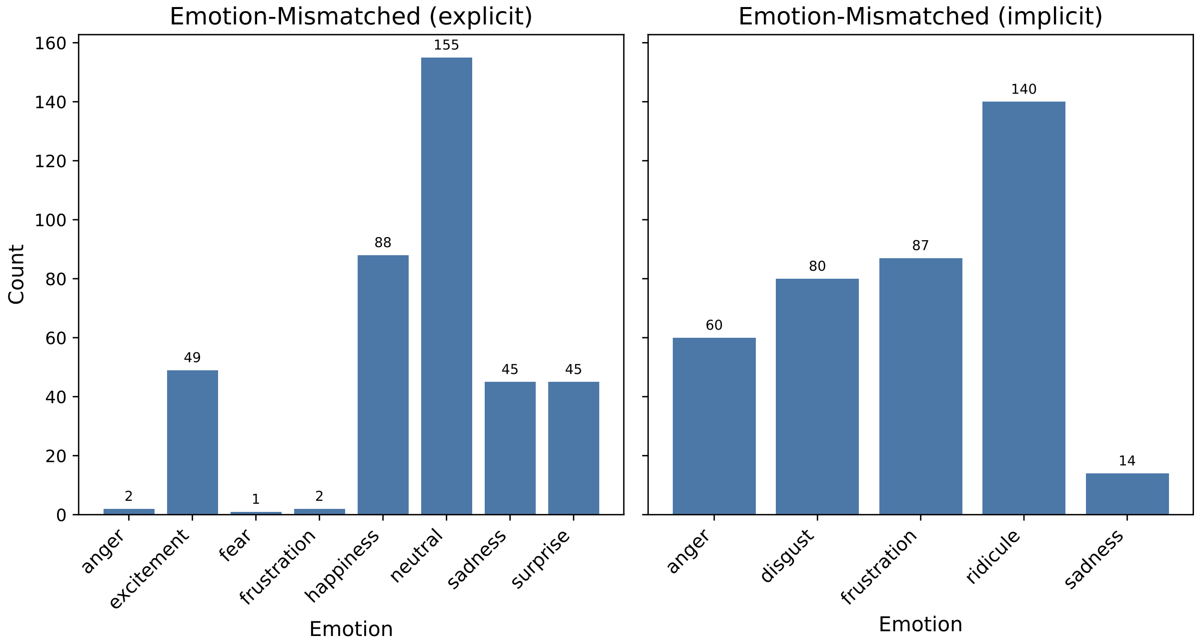


Figure 6. Ground-truth label distributions for Emotion-Mismatched (explicit) and Emotion-Mismatched (implicit).

A.4 Large Audio Language Models Evaluated in LISTEN

We evaluate six recent large audio–language models (LALMs) spanning both open-weight and proprietary systems, as summarized in Table 3. The open-weight group includes the Qwen and Baichuan series, which represent the strongest publicly released multilingual models with unified speech–text understanding capabilities. All models are evaluated in zero-shot settings without fine-tuning to ensure comparability across conditions.

A.5 Analysis of Coarse-Grained Emotion Label Mapping

To enable comparison with prior emotion recognition work and to analyze the effect of label granularity on model performance, we additionally report results using coarse-grained emotion labels. Specifically, we map the original fine-grained emotion annotations into five coarse emotion categories: sad, happy, neutral, surprise, and angry.

This mapping follows common practice in speech emotion recognition, where closely related affective states are merged into broader emotional classes. The mapping is defined as follows:

Angry: frustration, anger, angry, disgust, fear, ridicule

Table 3. Large audio–language models evaluated in the LISTEN benchmark. "?" indicates unspecified public information.

Model	**Identifier**	**Creator**	**Access Type**	**Release Date**	**Params**
Open-weight Audio–Language Models					
Qwen3-Instruct	Qwen3-4B-Instruct-2507	Alibaba Cloud	Open-weight	2025-08-06	4.02B
Qwen2.5-Omni (7B)	qwen2.5-omni-7b	Alibaba Cloud	Open-weight	2025-05-13	10.7B
Qwen3-Omni (30B)	qwen3-omni-30b	Alibaba Cloud	Open-weight	2025-09-026	35.3B
Baichuan-Omni (1.5)	baichuan-omni-1.5	Baichuan Inc.	Open-weight	2025-01-26	11B
Closed-weight Audio–Language Models					
Gemini 2.5 Flash	gemini-2.5-flash	Google DeepMind	API	2025-06	?
Gemini 2.5 Pro	gemini-2.5-pro	Google DeepMind	API	2025-06	?

Model	Neutral Text	Neutral Audio	Neutral Text+Audio	Matched Text	Matched Audio	Matched Text+Audio	Mismatched Text	Mismatched Audio	Mismatched Text+Audio	Paralinguistic
Baichuan-1	.31/.21	.16/.12	.15/.09	.24/.20	.20/.18	.24/.21	.23/.18	.23/.19	.25/.21	.37/.28
Gemini-2.5 Flash	.27/.19	.24/.21	.24/.22	.28/.21	.23/.20	.28/.24	.13/.13	.13/.08	.18/.10	.25/.14
Gemini-2.5 Pro	.97/.14	.34/.33	.40/.39	.29/.22	.28/.23	.32/.25	.13/.14	.24/.20	.28/.24	.23/.10
Qwen-2.5 Omni	.30/.22	.34/.33	.19/.16	.29/.21	.32/.25	.32/.25	.14/.14	.22/.09	.18/.08	.33/.17
Qwen-3 Instruct	.30/.18	–	–	.30/.21	–	–	.26/.24	–	–	–
Qwen-3 Omni	.33/.26	.33/.17	.24/.23	.29/.24	.33/.29	.29/.25	.16/.16	.19/.09	.18/.08	.30/.17

Table 4. UAR / Macro-F1 across models and experimental conditions.

Happy: happiness, happy, excitement, excited

Sad: sadness, sad

Surprise: surprise

Neutral: neutral, calm

Results are shown in Table 5. Across all evaluated models, we observe that improvements from fine-grained to coarse-grained evaluation are highly condition-dependent. In the Neutral-Text and Emotion-Matched conditions, accuracy increases are relatively modest. In contrast, the Emotion-Mismatched condition exhibits substantial performance gains.

This effect can be traced to the distribution of fine-grained labels within emotion-mismatched samples, which predominantly consist of sarcastic speech. In the original fine-grained annotation space, sarcastic utterances are spread across several closely related emotion categories, including frustration, anger, disgust, fear, and ridicule. When these categories are merged into the angry class, a large fraction of predictions that previously counted as errors become correct under the coarse-grained evaluation.

This result is aligned with confusion patterns shown in Figure 3. In both the audio-only and text+audio settings, we observe strong vertical concentration on ridicule and, to a lesser extent, frustration, indicating that models systematically assign emotionally conflicting samples to these categories regardless of the ground-truth fine-grained label. This pattern suggests that LALMs detect the presence of emotional conflict but resolve it by collapsing diverse sarcastic emotions into a small subset of dominant classes.

Taken together, this analysis demonstrates that the apparent performance gains under coarse-grained evaluation primarily reflect error absorption through label aggregation, rather than improved fine-grained emotional discrimination. As such, coarse-grained results should be interpreted as measuring high-level affect recognition, while fine-grained evaluation remains essential for diagnosing nuanced emotional understanding.

A.6 Detailed Cross-Modality Performance Analysis

Figure 7 compares detailed model performance across all modality–condition pairs in the LISTEN benchmark. Each axis represents a distinct evaluation setting, spanning Neutral-Text, Emotion-Matched, Emotion-Mismatched, and Paralinguistic conditions in both text and audio modalities. Across models, accuracy peaks in the Neutral-Text condition, where lexical content alone provides strong cues for emotion inference. Performance declines sharply in Audio-only and Paralinguistic settings, confirming that current large audio–language models (LALMs) struggle to extract affective meaning from acoustic cues. Even in Emotion-Matched scenarios, where lexical and prosodic signals align, the gains remain modest, suggesting limited multimodal integration. Among the evaluated systems,

Table 5. Accuracy (with prediction marginal distribution baseline in parenthesis—see subsection 4.1) are reported. **Bold** values highlight the highest value and underlined values highlight the second-highest value in each experiment. For each condition, the Average column represents the mean of the audio and text+audio settings. The Overall Average is computed as the mean accuracy across all audio and text+audio results from the four experimental conditions (seven modalities in total).

	Neutral-Text				**Emotion-Matched**				**Emotion-Mismatched**				**Paralinguistic**	**Overall**
Model	Text	Audio	Text+Audio	Average	Text	Audio	Text+Audio	Average	Text	Audio	Text+Audio	Average	Audio	Average
Uniform Guess	20.0	20.0	20.0	20.0	20.0	20.0	20.0	20.0	20.0	20.0	20.0	20.0	20.0	20.0
Majority Guess	100	16.9	16.9	16.9	26.5	26.5	26.5	26.5	39.0	39.0	39.0	39.0	32.6	28.2
Open-Weight Model														
Qwen3-instruct	70.8 (70.8)	–	–	–	40.0 (20.4)	–	–	–	90.5 (90.5)	–	–	–	–	–
Qwen2.5-Omni-7B	84.1 (84.1)	40.1 (21.3)	24.5 (18.1)	32.3	43.7 (22.5)	45.0 (22.4)	46.5 (22.8)	45.8	30.2 (17.0)	83.7 (83.3)	67.1 (67.7)	75.4	–	51.1
Qwen3-Omni-30B	84.0 (84.0)	35.0 (19.7)	30.1 (19.0)	32.5	46.0 (22.8)	49.9 (22.4)	50.3 (22.7)	50.1	30.8 (15.4)	75.7 (75.7)	71.2 (71.1)	73.5	–	52.0
Baichuan-Omni-1.5	80.2 (80.2)	21.1 (17.4)	19.1 (17.3)	20.1	43.6 (23.0)	40.5 (22.4)	45.4 (23.3)	43.0	87.0 (87.2)	**91.3** (91.2)	91.2 (91.5)	91.2	**40.8** (17.9)	49.9
Closed-Weight Model														
Gemini2.5-Flash	81.5 (81.5)	37.3 (23.0)	31.9 (20.0)	34.6	**59.0** (40.4)	**50.0** (25.1)	**54.3** (27.8)	52.1	26.6 (14.6)	50.7 (50.1)	59.1 (59.4)	54.9	–	47.2
Gemini2.5-Pro	**96.6** (96.6)	**51.0** (20.7)	**56.9** (20.8)	53.9	52.2 (26.1)	46.1 (23.5)	52.9 (24.5)	49.5	**92.5** (92.3)	89.5 (89.5)	**94.4** (94.0)	91.9	24.7 (15.5)	59.3

Gemini 2.5 Pro achieves the highest overall performance, followed by Qwen3-Omni-30B, yet all models display the same qualitative trend of lexical dominance.

A.7 Confusion Matrices For All Models

This presents the complete set of confusion matrices for all evaluated models across the three experimental conditions (Neutral-Text, Emotion-Matched, and Emotion-Mismatched) and three input modalities (Text, Audio, and Text+Audio). (see Figure 8, Figure 9, Figure 10, Figure 11) These figures provide a detailed view of each model's prediction distribution across emotion categories, complementing the main results discussed in section 5.

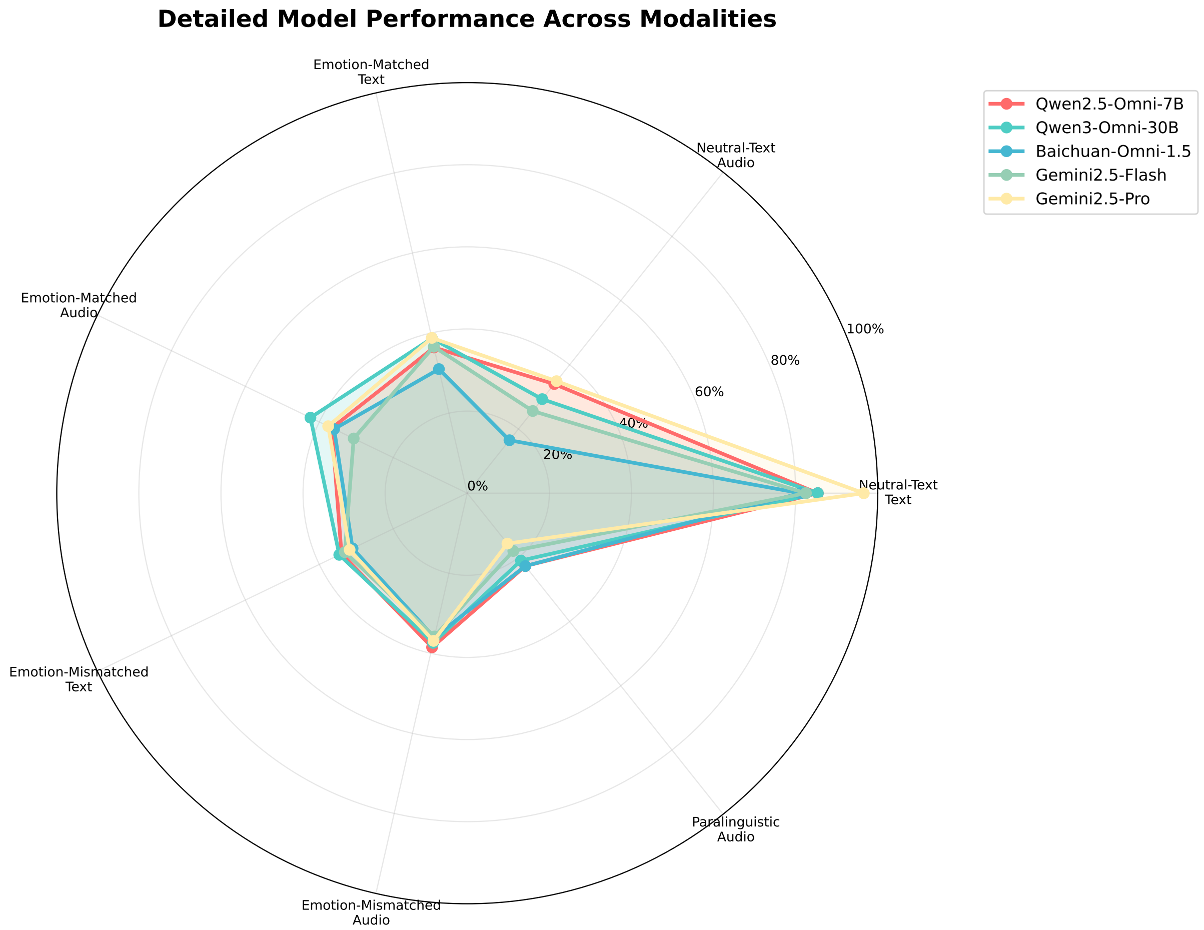


Figure 7. Radar plot comparing detailed model performance across all modality–condition pairs in the LISTEN benchmark. Each axis represents a specific evaluation setting (e.g., Neutral-Text, Emotion-Matched, Emotion-Mismatched, Paralinguistic) under text-only, audio-only, and text+audio modalities.

A.8 Condition Examples

We provides representative examples for each condition. (see Figure 14, Figure 15, Figure 16, Figure 17, Figure 18, Figure 19, Figure 20, and Figure 21 Figure 12 and Figure 13)

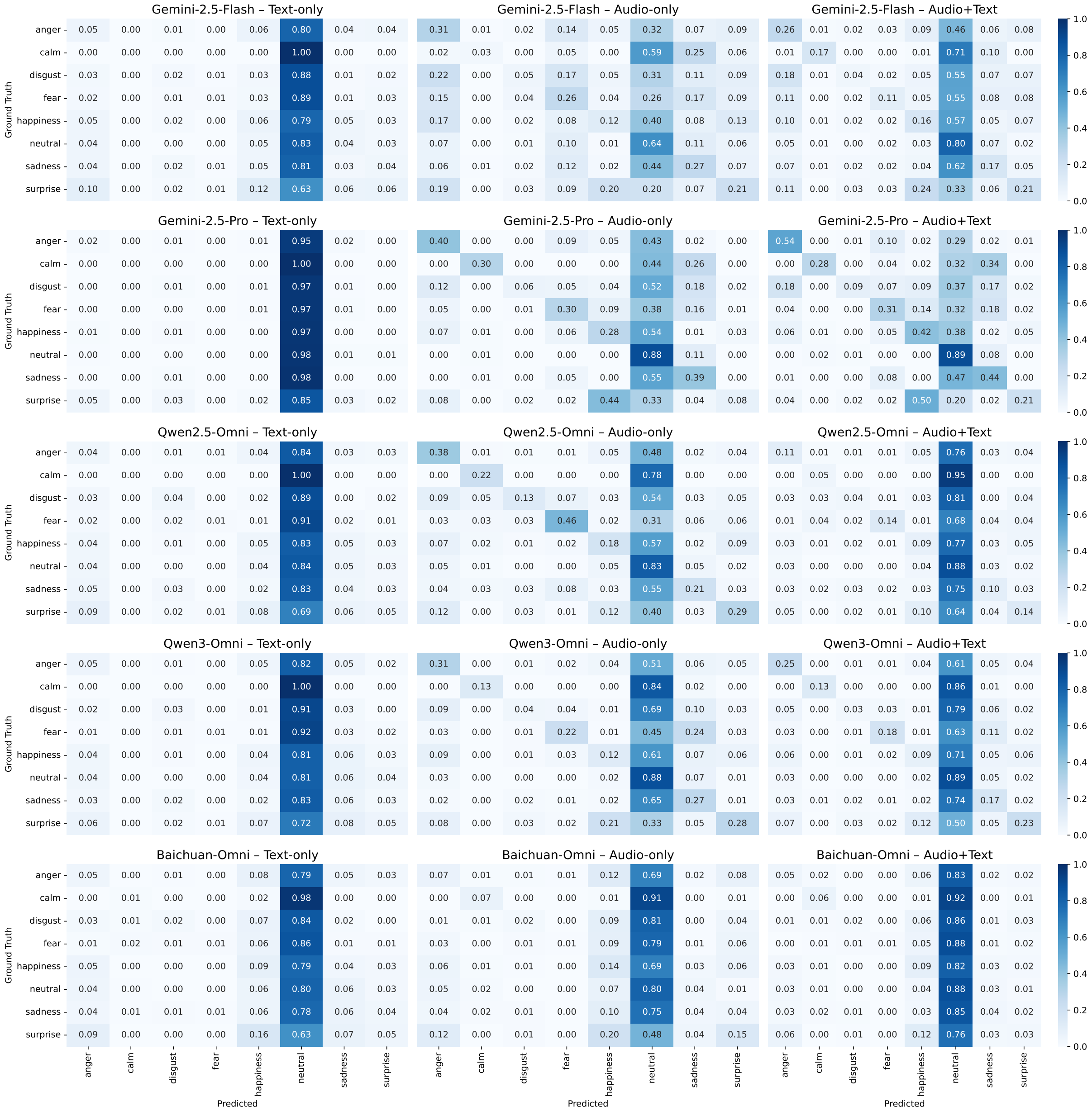


Figure 8. Confusion matrices showing all models' emotion recognition performance for Neutral-Text condition in the LISTEN benchmark. Row-normalized matrices display prediction distributions for each tested with text-only, audio-only, and audio+text modalities.

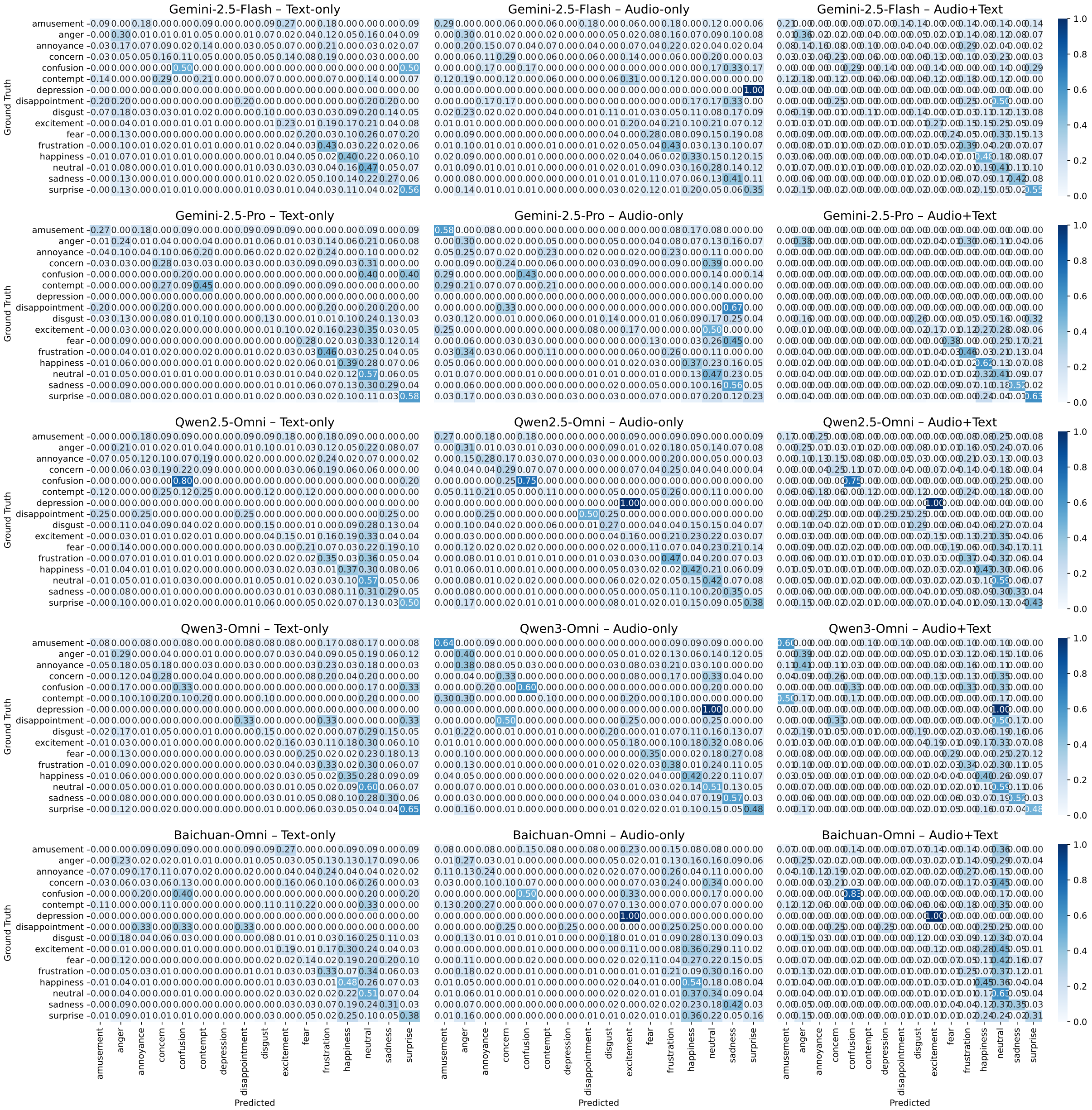


Figure 9. Confusion matrices showing all models' emotion recognition performance for Emotion-Matched condition in the LISTEN benchmark. Row-normalized matrices display prediction distributions for each tested with text-only, audio-only, and audio+text modalities.

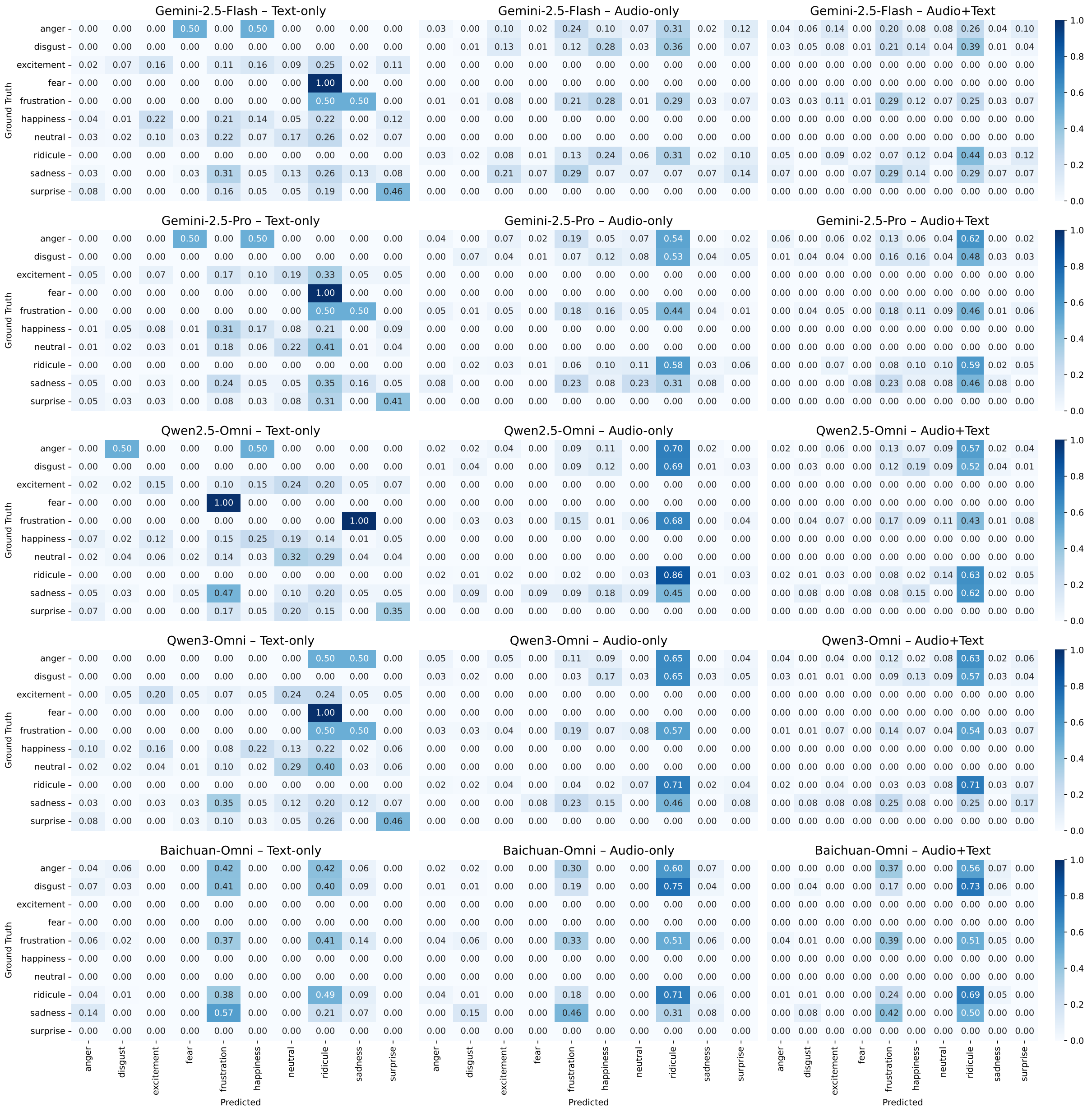


Figure 10. Confusion matrices showing all models' emotion recognition performance for Emotion-Mismatched condition in the LISTEN benchmark. Row-normalized matrices display prediction distributions for each tested with text-only, audio-only, and audio+text modalities.

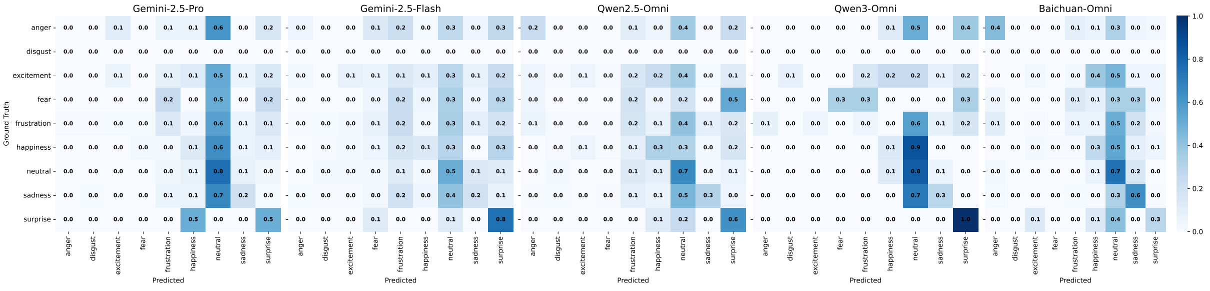


Figure 11. Confusion matrices showing all models' emotion recognition performance for Paralinguistic condition in the LISTEN benchmark. Row-normalized matrices display prediction distributions for each tested with audio-only modality.

Neutral-Text Text-only

Modality: Text-only

Sample ID: `SAMPLE_7c8b53fb`

Transcription: *"Kids are talking by the door."*

Prompt: Read the transcription and classify the emotion. Based on the content of this text, what emotion would the person likely be feeling?

A. anger
B. fear
C. disgust
D. neutral
E. sadness
F. surprise
G. calm
H. happiness

Expected Response: D (neutral)

Model Prediction: D (neutral)

Ground Truth: neutral

Evaluation: Correct

Figure 12. Example of a Neutral-Text text only entry in the LISTEN benchmark. The model correctly identifies the statement's emotion (*neutral*) when both lexical cues is available.

Neutral-Text Audio-only

Modality: Audio-only

Sample ID: `SAMPLE_7c8b53fb`

Audio: RAVDESS_train_0333

Prompt: Listen to the audio and classify the emotion. What emotion does the speaker convey through their tone?

A. surprise
B. sadness
C. fear
D. anger
E. calm
F. happiness
G. neutral
H. disgust

Expected Response: D (anger)

Model Prediction: D (anger)

Ground Truth: anger

Evaluation: Correct

Figure 13. Example of a Neutral-Text audio-only entry in the LISTEN benchmark. The model correctly infers the intended emotion (*anger*) based on vocal prosody alone.

Neutral-Text Text+Audio

Modality: Text+Audio

Sample ID: `SAMPLE_7c8b53fb`

Transcription: *"Kids are talking by the door."*

Audio: RAVDESS_train_0333

Prompt: Listen to the audio and read the transcription, then classify the emotion. What emotion does the speaker convey through their tone?

A. surprise
B. sadness
C. fear
D. anger
E. calm
F. happiness
G. neutral
H. disgust

Expected Response: D (anger)

Model Prediction: D (anger)

Ground Truth: anger

Evaluation: Correct

Figure 14. Example of a Neutral-Text text+audio entry in the LISTEN benchmark. The model correctly identifies the speaker's emotion (*anger*) when both lexical and prosodic cues are available.

Emotion-Matched Text-only

Modality: Text-only

Sample ID: `SAMPLE_9b76ea7d`

Transcription: *"What the hell is this?"*

Prompt: Read the transcription and classify the emotion. From the semantic content alone, what emotion is being expressed?

A. neutral
B. sadness
C. excitement
D. frustration
E. fear
F. disgust
G. happiness
H. anger
I. surprise

Expected Response: H (frustration)

Model Prediction: D (neutral)

Ground Truth: frustration

Evaluation: Incorrect

Figure 15. Example of an Emotion-Matched text-only entry from the LISTEN benchmark. The model misclassified an explicitly frustration utterance (*"What the hell is this?"*) as *neutral*, illustrating overgeneralization across semantically related negative emotions.

Emotion-Matched Audio-only

Modality: Audio-only

Sample ID: `SAMPLE_dd0f6e9d`

Audio: IEMOCAP_Session5_Ses05M_script01_1b_F030

Prompt: Listen to the audio and classify the emotion. Based on the vocal expression, what emotion is the speaker feeling?

A. frustration
B. anger
C. neutral
D. excitement
E. happiness
F. surprise
G. disgust
H. fear
I. sadness

Expected Response: A (frustration)

Model Prediction: A (frustration)

Ground Truth: frustration

Evaluation: Correct

Figure 16. Example of an Emotion-Matched audio-only entry from the LISTEN benchmark. The model correctly interprets prosodic cues to identify the emotion as *frustration*, showing sensitivity to vocal intensity and tone even without textual input.

Emotion-Matched Text+Audio

Modality: Text+Audio

Sample ID: `SAMPLE_dd0f6e9d`

Audio: IEMOCAP_Session5_Ses05M_script01_1b_F030

Transcription: *"What the hell is this?"*

Prompt: Listen to the audio and read the transcription, then classify the emotion. Based on the vocal expression, what emotion is the speaker feeling?

A. frustration
B. anger
C. neutral
D. excitement
E. happiness
F. surprise
G. disgust
H. fear
I. sadness

Expected Response: A (frustration)

Model Prediction: A (frustration)

Ground Truth: frustration

Evaluation: Correct

Figure 17. Example of an Emotion-Matched text+audio entry from the LISTEN benchmark. The model correctly identifies *frustration* when integrating both lexical and prosodic cues, demonstrating effective multimodal fusion under congruent emotional alignment.

Emotion-Mismatched Text-only

Modality: Text-only

Sample ID: `SAMPLE_955399e0`

Transcription: *"You're right, the party's fantastic. Please, tell me more. I haven't heard enough about it all week because hearing about that never gets old!"*

Prompt: Read the transcription and classify the emotion. What emotion is conveyed by the words in this statement?

A. surprise
B. excitement
C. sadness
D. disgust
E. fear
F. neutral
G. anger
H. happiness
I. frustration
J. ridicule

Expected Response: B (excitement)

Model Prediction: B (excitement)

Ground Truth: excitement (explicit emotion label)

Evaluation: Correct

Figure 18. Example of an Emotion-Mismatched text-only entry from the LISTEN benchmark. Although the lexical content expresses *excitement*, the corresponding audio (not shown) conveys *ridicule*, highlighting the designed lexical–prosodic conflict characteristic of this condition.

Emotion-Mismatched Audio-only

Modality: Audio-only

Sample ID: `SAMPLE_c52e71d0`

Audio: MUStARD_PRO_1_7575_u_3B

Prompt: Listen to the audio and classify the emotion. What emotion is communicated through the speaker's vocal prosody?

A. disgust
B. neutral
C. ridicule
D. frustration
E. sadness
F. anger
G. excitement
H. fear
I. surprise
J. happiness

Expected Response: F (anger)

Model Prediction: G (excitement)

Ground Truth: anger (implicit emotion label)

Evaluation: Incorrect

Figure 19. Example of an Emotion-Mismatched audio-only entry from the LISTEN benchmark. The lexical content is superficially positive (*"You're right, the party's fantastic"*), but the prosody expresses irritation and *anger*. The model incorrectly predicts *excitement*, indicating difficulty in resolving sarcastic or contrastive vocal tone.

Emotion-Mismatched Text+Audio

Modality: Text+Audio

Sample ID: `SAMPLE_c52e71d0`

Audio: `MUStARD_PRO_1_7575_u_3B`

Transcription: *"You're right, the party's fantastic. Please, tell me more. I haven't heard enough about it all week because hearing about that never gets old!"*

Prompt: Listen to the audio and read the transcription, then classify the emotion. What emotion is communicated through the speaker's vocal prosody?

A. neutral
B. disgust
C. anger
D. sadness
E. excitement
F. fear
G. ridicule
H. frustration
I. happiness
J. surprise

Expected Response: C (anger)

Model Prediction: E (excitement)

Ground Truth: anger (implicit emotion label)

Evaluation: Incorrect

Figure 20. Example of an Emotion-Mismatched text+audio entry from the LISTEN benchmark. Despite access to both modalities, the model incorrectly predicts *excitement* instead of the intended *anger*, suggesting overreliance on lexical positivity rather than prosodic dissonance—a hallmark challenge in sarcasm and irony understanding.

Paralinguistic Audio-only

Modality: Audio-only

Sample ID: `SAMPLE_54df39ff`

Audio: `IEMOCAP_Session5_Ses05F_impro03_F006`

Prompt: Listen to the audio and classify the emotion. What emotional tone is conveyed by the literal meaning of this statement?

A. anger
B. happiness
C. fear
D. sadness
E. surprise
F. frustration
G. excitement
H. disgust
I. neutral

Expected Response: G (excitement)

Model Prediction: B (happiness)

Ground Truth: excitement

Evaluation: Incorrect

Figure 21. Example of a Paralinguistic audio-only entry from the LISTEN benchmark. The utterance contains only nonverbal laughter, labeled as *excitement*. The model incorrectly classifies it as *happiness*, revealing challenges in distinguishing subtle affective intent from nonverbal vocalizations.