When Audio and Text Disagree: Benchmarking Text Bias in Large Audio-Language Models under Cross-Modal Inconsistencies

Cheng Wang[†] **Gelei Deng**[‡*] **Xianglin Yang**[†] **Han Qiu**[§] **Tianwei Zhang**[‡]
[†] National University of Singapore
[‡] Nanyang Technological University
[§] Tsinghua University
wangcheng@u.nus.edu

Abstract

Large Audio-Language Models (LALMs) are enhanced with audio perception capabilities, enabling them to effectively process and understand multimodal inputs that combine audio and text. However, their performance in handling conflicting information between audio and text modalities remains largely unexamined. This paper introduces MCR-BENCH, the first comprehensive benchmark specifically designed to evaluate how LALMs prioritize information when presented with inconsistent audio-text pairs. Through extensive evaluation across diverse audio understanding tasks, we reveal a concerning phenomenon: when inconsistencies exist between modalities, LALMs display a significant bias toward textual input, frequently disregarding audio evidence. This tendency leads to substantial performance degradation in audio-centric tasks and raises important reliability concerns for real-world applications. We further investigate the influencing factors of text bias, and explore mitigation strategies through supervised finetuning, and analyze model confidence patterns that reveal persistent overconfidence even with contradictory inputs. These findings underscore the need for improved modality balance during training and more sophisticated fusion mechanisms to enhance the robustness when handling conflicting multi-modal inputs[1].

1 Introduction

With the rise of Large Audio-Language Models (LALMs) (Chu et al., 2024; Tang et al., 2023; Gong et al., 2023), there has been significant progress in developing applications and systems capable of processing both auditory and textual information for complex tasks. These models, often built upon Large Language Models (LLMs) with specialized audio encoders, have demonstrated impressive capabilities in various audio-centric tasks including Audio Question Answering (Lipping et al., 2022), Sound Event Detection (Mesaros et al., 2021), and Speech Recognition (Radford et al., 2022). The wide deployment of LALMs across various domains reflects their growing importance in bridging human auditory experience with machine intelligence.

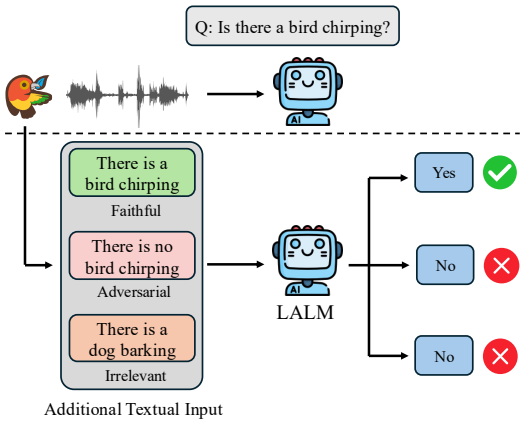

Figure 1: Illustration of LALMs handling users' input with conflicts across the text and audio modalities.

To facilitate the development of LALMs, numerous benchmarks and datasets have been established for performance evaluations (Wang et al., 2025; Yang et al., 2024). However, they typically assume harmonious or complementary relationships between audio and text inputs. In particular, standard datasets often pair audio samples with accurate textual descriptions or questions that precisely align with the audio content. This idealized evaluation approach, while useful for basic capability assessment, fails to capture the *robustness* of these models in handling real-world scenarios where the input of different modalities contains conflicting information. Researchers have demonstrated that the inconsistent inputs could significantly degrade the performance of LLMs (Shi et al., 2023; Liu et al., 2024) or Large Vision-Language Models

* Corresponding Author
[1]The project is available at https://github.com/WangCheng0116/MCR-BENCH

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 4878–4888
November 4-9, 2025

(LVLMs) (Liu et al., 2025b; Deng et al., 2025). However, there is still a lack of systematic investigation into how LALMs behave when faced with contradictory inputs, representing a significant gap in our understanding of these models' reliability.

The above research gap drives the motivation of our study, where we aim to systematically evaluate and mitigate the limitations of contemporary LALMs under conflicting modal information. We believe this is crucial for ensuring their safe and dependable use in real-world applications. We hypothesize that when faced with inconsistent audio and text inputs, LALMs may exhibit a bias toward one modality—either audio or text—over the other, potentially leading to suboptimal performance in audio-centric tasks. This preferential behavior could undermine the models' ability to effectively integrate and reconcile multi-modal data, which is essential for their robustness in complex, dynamic environments.

To validate our hypothesis, we introduce MCR-BENCH, a comprehensive Modal Conflict Resolution Benchmark for LALMs. Departing from traditional clean audio-text pairs, MCR-BENCH comprises 3,000 specially constructed samples across three audio-centric tasks, where each audio input is systematically paired with adversarial, faithful, and irrelevant textual descriptions. Through extensive experiments evaluating six state-of-the-art LALMs on MCR-BENCH, we reveal a consistent and substantial preference for textual input over audio, leading to severe performance degradation in the presence of misleading text. This modality bias is evident across diverse tasks and model architectures, indicating a widespread issue in current LALM designs.

Beyond characterizing textual bias, we further explore mitigation strategies and analyze internal LLM state differences when processing clean versus contradictory samples. We find that simple prompting techniques—such as bias-aware or audio-prioritized instructions—yield only limited improvements, while supervised finetuning on conflict-rich data offers more promising, though still incomplete, mitigation. Further analysis of model behavior reveals that LALMs remain highly confident even when relying on contradictory textual information, and internal representation studies suggest they internally detect cross-modal inconsistencies without appropriately modulating their outputs. These findings underscore a disconnect between latent awareness and output reasoning, highlighting the need for architectural and training-level innovations to achieve truly robust multi-modal reasoning in audio-language models. Importantly, these insights illuminate a promising path forward: leveraging mechanism interpretation to develop new solutions for robust audio-language models.

2 Related Work

LALMs Performance Benchmarking. Large Audio-Language Models (LALMs) have recently gained significant attention for their ability to process audio inputs and generate textual responses. Researchers have established task-specific benchmarks for audio understanding capabilities, such as AudioBench (Wang et al., 2025) and AIR-Bench (Yang et al., 2024). These benchmarks predominantly assume aligned or complementary audio-text relationships, leaving the models' behavior under conditions of modal conflict largely unexplored. While recent work has begun addressing evaluation comprehensiveness, the assumption of modal harmony persists, creating a critical gap in our understanding of LALMs' reliability in real-world scenarios where inputs across modalities may contain inconsistencies.

Robustness of LALMs. Prior research has focused on two primary dimensions of audio models' robustness: vulnerability to adversarial attacks and resilience against natural perturbations. While Carlini and Wagner (2018) and Qin et al. (2019) demonstrated concerning susceptibilities to targeted and imperceptible adversarial examples, defensive strategies such as data augmentation techniques proposed by Park et al. (2019) and self-supervised learning frameworks from Baevski et al. (2020) have shown promise in improving model resilience. Despite these advances, the field requires more systematic evaluations and comprehensive frameworks to address the multifaceted challenges of real-world audio processing.

Distraction in Inputs. Recent studies highlight the challenge of distraction in input processing across both language and multimodal models. For LLMs, Huang et al. (2025) introduced Contextual Distraction Vulnerability, demonstrating how irrelevant but semantically coherent context significantly degrades model performance. To address this challenge, retrieval-augmented contrastive learning approaches have been explored to enhance focus on relevant information in long-context tasks (Wu

et al., 2024). The distraction problem extends to multimodal systems as well, with Deng et al. (2025) and Liu et al. (2025b) systematically analyzing how Vision-Language Models exhibit substantial performance degradation when confronted with conflicting visual and textual inputs. These studies establish that inconsistent or distracting information across modalities presents a fundamental challenge for robust AI systems. Our work extends this line of inquiry to the audio domain, investigating how LALMs prioritize information when faced with similar cross-modal inconsistencies.

3 MCR-BENCH

We introduce MCR-BENCH, a benchmark specifically designed to evaluate how LALMs process and reconcile conflicting audio-text inputs. For each audio sample in our benchmark, we systematically construct three types of textual contexts:

- **Faithful:** Accurate descriptions that correctly represent the audio content.
- **Adversarial:** Deliberately misleading descriptions that contradict the audio content.
- **Irrelevant:** Semantically unrelated descriptions that have minimal topical overlap with the audio content.

These variations allow us to systematically evaluate LALMs' ability to prioritize relevant audio information, resist misleading textual cues, and maintain robust performance when faced with conflicting or irrelevant cross-modal inputs. Below we elaborate how these three types of text variations are constructed.

3.1 Data Sources

MCR-BENCH covers three different types of audio understanding tasks (sound question answering, speech emotion recognition, and vocal sound classification) to ensure a comprehensive evaluation across diverse audio domains. It is extensible for supporting other audio-text tasks as well.

- **Audio Question Answering (AQA):** We utilize ClothoAQA (Lipping et al., 2022), a dataset comprising 1,991 audio samples from the Clotho (Drossos et al., 2019) dataset, each paired with six crowdsourced questions and corresponding answers, totaling 35,838 question-answer pairs. This component evaluates natural language understanding of general audio content.
- **Speech Emotion Recognition (SER):** We incorporate MELD (Poria et al., 2019), a multimodal multi-party dataset containing over 1,400 dialogues and 13,000 utterances from the TV series Friends, annotated with seven emotion labels and sentiment. This tests the model performance on human speech with emotional content.
- **Vocal Sound Classification (VSCn):** We include VocalSound (Gong et al., 2022), which features non-verbal human vocalizations across different acoustic conditions, challenging models to recognize human vocal sounds beyond speech.

3.2 Text Variation Construction

To generate systematic variations in textual contexts, we create three distinct textual conditions for each audio sample:

- **Faithful Text Generation:** We employ GPT-4o (OpenAI, 2024) with one-shot learning to create factual statements that accurately represent the audio content based on original question-answer pairs.
- **Adversarial Text Generation:** Using the same GPT-4o framework, we generate non-factual statements that directly contradict the audio content. Appendix A shows the prompt template used for this adversarial generation process.
- **Irrelevant Text Selection:** We select irrelevant textual descriptions based on sentence similarity calculations between the true caption and all captions from AudioCaps (Kim et al., 2019). We choose descriptions with minimal semantic overlap while maintaining plausible text structure.

3.3 Evaluation Metrics

To quantify modal conflict resolution capabilities of different LALMs, we define N as the total number of samples, C_{neutral} as the number of correct predictions under neutral conditions (where only audio input is provided without any textual description), and C_t as the number of correct predictions with text condition $t \in \{neu, fth, adv, irr\}$ for faithful, adversarial, and irrelevant conditions respectively. For evaluation, we use a prompt template shown in Figure 2 that instructs models to answer questions while being aware that the provided textual descriptions may contain inaccuracies. Specifically, we use the following metrics.

Accuracy. For each textual description type, we

calculate the accuracy as:

$$\text{Acc}_t = \frac{C_t}{N}.$$

Normalized Accuracy. This metric measures how the model is affected by different types of textual input. It can be expressed as:

$$\text{Norm}_t = \frac{C_t}{C_{\text{neu}}}.$$

Macro Accuracy. This metric is defined as the average accuracy of three different types:

$$\text{Macro} = \frac{\text{Acc}_{fth} + \text{Acc}_{adv} + \text{Acc}_{irr}}{3}.$$

Text Influence Rate (TIR). TIR quantifies how much the textual input influences the model's predictions across different text conditions. We define Δ^t_{ic} as the number of samples with prediction changes from incorrect to correct and Δ^t_{ci} as those changing from correct to incorrect under text condition t. TIR for each text condition is:

$$\text{TIR}_t = \begin{cases} \frac{\Delta^{fth}_{\text{ic}}}{N} & \text{if } t = fth \\ \frac{\Delta^{adv}_{\text{ci}}}{N} & \text{if } t = adv \\ \frac{\Delta^{irr}_{\text{ic}} + \Delta^{irr}_{\text{ci}}}{N} & \text{if } t = irr \end{cases}$$

It evaluates how the model utilizes faithful information, opposing misleading information and ignoring irrelevant information.

Modal Robustness Score (MRS). This metric quantifies the resistance to potentially misleading textual information, indicating how well a model maintains audio-based performance despite contradictory or irrelevant textual inputs:

$$\text{MRS} = \alpha \cdot \frac{\text{Acc}_{adv}}{\text{Acc}_{neu}} + (1 - \alpha) \cdot \frac{\text{Acc}_{irr}}{\text{Acc}_{neu}}$$

where $\alpha = 0.8$ emphasizes adversarial robustness, as most models exhibit near-zero Acc_{adv} values.

Prompts for Evaluation

The following text provides descriptions derived from audio content. This information may be accurate, inaccurate, or irrelevant. Answer the question with discretion regarding how you use these details.
Description: <DESCRIPTION>
Question: <QUESTION>

Figure 2: The prompt used for evaluating LALMs' response to potentially inconsistent audio-text pairs.

4 Benchmarking Text Bias

Our evaluation encompasses a diverse range of state-of-the-art LALMs: Qwen-Audio-Chat (Chu et al., 2023), Qwen2-Audio-Instruct (Chu et al., 2024), Gazelle (AI, 2024), SALMONNN-7B and SALMONNN-13B (Tang et al., 2023), Audio-Flamingo2 (Ghosh et al., 2025) and SeaLLMs-Audio-7B (Liu et al., 2025a).

4.1 Main Results

Experimental results are summarized in Table 1. We observe strong text bias across all models. LALMs consistently prioritize textual information over audio evidence when faced with contradictions between modalities, regardless of their model architecture or underlying training methodology. When provided with adversarial textual descriptions that contradict audio content, all models exhibit dramatic performance drops. For instance, on the Audio Question Answering task, accuracies drops from 87.8% to 1.7% for Qwen-Audio-Chat and from 87.5% to 1.5% for Qwen2-Audio-Instruct—representing over 98% performance deterioration. Even more strikingly, on the Speech Emotion Recognition task, four of the seven tested models show complete susceptibility to adversarial text, with accuracy dropping to precisely 0.0%. This pattern holds across all datasets, with TIR consistently above 95% for most models, clearly demonstrating that these systems overwhelmingly favor textual inputs when resolving cross-modal conflicts.

4.2 Comparisons Across Models

While text bias is universal across all tested models, some demonstrate notably higher resilience to misleading textual inputs than others. Audio-Flamingo2 stands out with substantially stronger modal robustness compared to other models, achieving significantly higher adversarial accuracy on Audio Question Answering task (35.3% versus below 3.5% for most competitors) and maintaining an MRS of 58.4%. Similarly, on the Speech Emotion Recognition task, Audio-Flamingo2 maintains 15.9% accuracy under adversarial conditions while most other models drop to near zero. SALMONN models also demonstrate relatively better resilience on Vocal Sound Classification, with the 7B and 13B versions maintaining 25.1% and 24.4% accuracy respectively under adversarial conditions, compared to 3.0% of Qwen-Audio-Chat. These quantitative

Benchmark Task	Model	Neutral	Faithful			Adversarial			Irrelevant			Macro	MRS
			Accuracy ↑	Norm ↑	TIR ↑	Accuracy ↑	Norm ↑	TIR ↓	Accuracy ↑	Norm ↑	TIR ↓		
AQA	Qwen-Audio-Chat	87.8	100.0	113.9	100.0	1.7	1.9	98.3	87.9	100.1	12.7	63.2	21.5
	Qwen2-Audio-Instruct	87.5	100.0	114.3	100.0	1.5	1.7	98.3	75.5	86.3	27.0	59.0	18.6
	SALMONN-7B	62.2	99.4	159.8	98.7	1.7	2.7	97.3	73.8	118.6	26.0	58.3	25.9
	SALMONN-13B	70.0	99.4	142.0	98.3	2.7	3.9	96.6	55.3	79.0	62.1	52.5	18.9
	Gazelle	60.5	87.2	144.1	86.1	3.5	5.8	96.4	43.6	72.1	53.7	44.8	19.1
	Audio-Flamingo2	68.0	90.4	132.9	82.5	35.3	51.9	58.7	57.5	84.6	38.7	61.1	58.4
	SeaLLMs-Audio-7B	72.8	99.9	137.2	99.6	1.3	1.8	98.4	81.8	112.4	15.6	61.0	23.9
VSC	Qwen-Audio-Chat	60.1	79.5	132.3	51.9	3.0	5.0	96.7	45.3	75.4	15.7	42.6	19.1
	Qwen2-Audio-Instruct	85.4	99.8	116.9	98.6	11.8	13.8	86.2	85.7	100.4	9.9	65.8	31.1
	SALMONN-7B	60.5	89.6	148.1	73.7	25.1	41.5	59.0	61.4	101.5	4.5	58.7	53.5
	SALMONN-13B	48.8	65.3	133.8	38.7	24.4	50.0	52.5	42.4	86.9	12.6	44.0	57.4
	Gazelle	18.2	100.0	549.5	100.0	0.0	0.0	100.0	16.9	92.9	15.3	39.0	18.6
	Audio-Flamingo2	30.0	98.8	329.3	98.7	1.3	4.3	97.3	25.3	84.3	26.7	41.8	20.3
	SeaLLMs-Audio-7B	65.2	98.4	150.9	95.4	7.1	10.9	88.3	49.7	76.2	18.7	51.7	24.0
SER	Qwen-Audio-Chat	24.5	99.9	407.8	99.9	0.1	0.4	99.6	14.8	60.4	25.9	38.3	12.4
	Qwen2-Audio-Instruct	41.8	100.0	239.2	100.0	0.0	0.0	100.0	27.8	66.5	39.4	42.6	13.3
	SALMONN-7B	25.1	98.7	393.2	98.3	0.1	0.4	99.6	36.4	145.0	36.1	45.1	29.3
	SALMONN-13B	46.9	100.0	213.2	100.0	0.0	0.0	100.0	45.3	96.6	5.0	48.4	19.3
	Gazelle	44.9	97.4	216.9	95.6	0.0	0.0	100.0	43.8	97.6	7.3	47.1	19.5
	Audio-Flamingo2	30.8	80.2	260.4	76.0	15.9	51.6	72.7	32.9	106.8	31.1	43.0	62.6
	SeaLLMs-Audio-7B	49.9	99.9	200.2	99.8	0.1	0.2	99.8	47.2	94.6	18.3	49.1	19.1

Table 1: **Performance comparison (%) of various LALMs on MCR-BENCH**. Results show accuracy and Text Influence Rate (TIR) across neutral, faithful, adversarial, and irrelevant text inputs. Darker background color indicate higher value.

differences suggest meaningful variations in how different architectures integrate and prioritize cross-modal information, though even the most robust models still show considerable vulnerability to text bias.

To investigate the relationship between parameter count and cross-modal robustness, we evaluated Audio-Flamingo2 (Ghosh et al., 2025) at three different scales (0.5B, 1.5B, and 3B) as detailed in Table 2. Our analysis reveals a consistent performance improvement as model size increases, with the largest 3B variant showing enhanced capabilities in both leveraging helpful textual information and resisting misleading inputs. However, the relatively modest gains in adversarial resistance compared to the significant parameter increase suggest that architectural innovations, rather than simple scaling, may be necessary to effectively address cross-modal conflicts.

4.3 Impact of Tasks and Text Relevance

The severity of text bias varies significantly across different audio understanding tasks, revealing a relationship between task complexity and susceptibility to misleading text. LALMs show particularly high vulnerability on emotion recognition tasks, where average adversarial accuracy across all models is just 2.3%, compared to 6.7% on Audio Question Answering task and 10.4% on Vocal Sound Classification task. Similarly striking is how irrelevant text affects performance differently across tasks—on Audio Question Answering, SeaLLMs-Audio-7B achieves 112.4% normalized accuracy with irrelevant text (improved performance), while on Speech Emotion Recognition task, SALMONN-7B reaches 145.0% of its neutral performance with irrelevant text. This variability in responses to different types of textual interference suggests that the interplay between audio and text processing is highly task-dependent, with semantically complex tasks showing different vulnerability patterns than more straightforward classification tasks.

To investigate how textual relevance affects model behavior, we quantify the semantic distance between textual descriptions and audio content, dividing samples into five bins from lowest to highest relevance. Using sentence embeddings to compute cosine similarity between text and audio captions, we evaluate performance across these relevance levels. As shown in Figure 3, surprisingly, there is no clear correlation between text relevance and the model's susceptibility to textual bias. The Text Influence Rate remains consistently high across all relevance bins for adversarial text, suggesting that LALMs' text bias persists regardless of semantic distance between modalities.

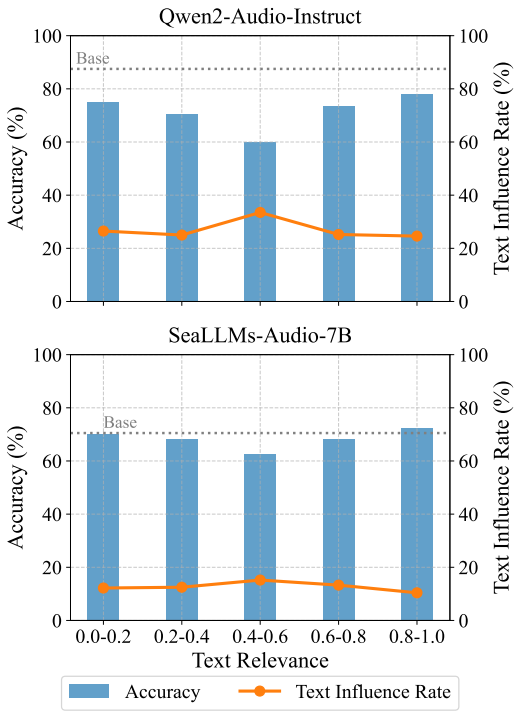

Figure 3: **Analysis of Text Relevance Impact.** Performance across five text relevance bins from lowest to highest. Blue bars (left axis) show accuracy under adversarial text conditions, while the orange line (right axis) represents the Text Influence Rate.

Size	Text Influence Rate			Macro ↑	MRS ↑
	Faith. ↑	Adv. ↓	Irr. ↓		
0.5B	75.36	66.38	44.20	54.60	58.10
1.5B	72.67	59.30	43.80	55.17	59.44
3B	**82.50**	**58.68**	**38.70**	**61.07**	**60.68**

Table 2: **The Effect of Model Sizes.** We experiment with Audio-Flamingo2 at three different parameter scales on ClothoAQA.

5 Understanding Text Bias

We perform in-depth analysis to disclose the causes of text bias in LALMs.

5.1 Confidence Analysis

To investigate whether LALMs exhibit appropriate uncertainty when faced with inconsistent inputs, we analyze confidence patterns in Qwen2-Audio-Instruct and SeaLLMs-Audio-7B across different textual conditions. For each prediction, we extract the maximum token probability as a confidence score, allowing us to quantify model certainty under modal conflict.

As shown in Figure 4, LALMs maintain remarkably high confidence scores even when processing adversarial textual inputs that contradict audio evidence. Surprisingly, confidence under adversarial conditions is comparable to or even higher than under faithful conditions, despite the dramatic performance degradation observed in our earlier experiments. Only with irrelevant text do we observe a slight reduction in confidence, though this decrease remains disproportionately small relative to performance impact. This overconfidence when making incorrect predictions indicates that LALMs not only prioritize text over audio but also do so with high certainty, suggesting these models lack effective calibration mechanisms to detect and appropriately respond to cross-modal inconsistencies.

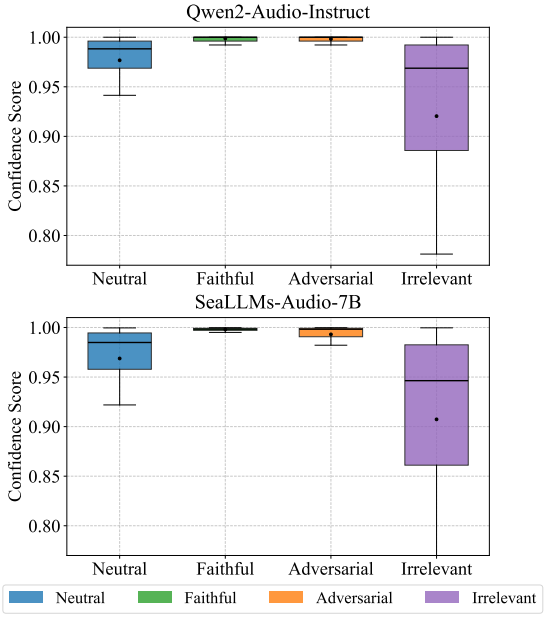

Figure 4: **Confidence Analysis Under Different Textual Conditions.** LALMs maintain high confidence scores across text conditions despite performance degradation with adversarial inputs.

5.2 Spectral Analysis

We analyze the intrinsic dimensionality of hidden representations when processing consistent versus inconsistent audio-text pairs. For N samples, we extract the last layer hidden states of the final token, resulting in two matrices: $A \in \mathbb{R}^{N\times d}$ from adversarial inputs and $F \in \mathbb{R}^{N\times d}$ from faithful inputs, where d represents the hidden state dimension. After centralizing these matrices, we perform Singular Value Decomposition (SVD):

$$A = U_A \Sigma_A V_A^T, \quad F = U_F \Sigma_F V_F^T$$

where $U_A, U_F \in \mathbb{R}^{N\times N}$ and $V_A, V_F \in \mathbb{R}^{d\times d}$ are orthogonal matrices.

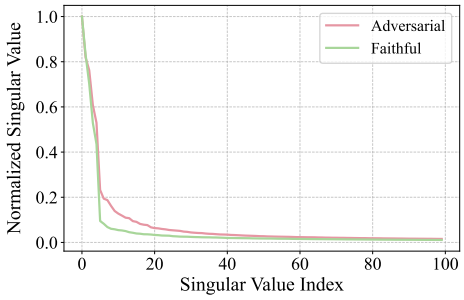

Figure 5: **Spectral analysis of hidden representations.** Normalized singular values for adversarial and faithful inputs from Qwen2-Audio-Instruct on subset of MCR-BENCH.

Using Qwen2-Audio-Instruct with the Vocal Sound Classification subset of MCR-BENCH, we plot the normalized singular values in Figure 5. The results reveal a rapid decay in singular values for both conditions, indicating that the model's representations lie in remarkably low-dimensional subspaces. Specifically, only 6 dimensions are needed to explain 95% of the variance in adversarial representations, while faithful representations require just 5 dimensions. This suggests that despite the high-dimensional embedding space, the model encodes audio-text information in compact, low-dimensional manifolds.

5.3 Separability Analysis

Building on our spectral analysis findings, we further investigate the separability between these low-dimensional subspaces. If the model internally distinguishes between faithful and adversarial inputs—despite producing confident yet incorrect outputs for adversarial cases—these subspaces should be linearly separable. We implement a 3:1 train-test split on the hidden representations from different model layers and train SVM and Random Forest classifiers to quantify this separability.

Table 3 presents the classification performance across different layers. The high accuracy (up to 98.0% with Random Forest at layer 32) confirms that these representation subspaces are highly separable, with the separation becoming more pronounced in deeper layers. This indicates that LALMs internally recognize inconsistencies between audio and text modalities, yet this awareness fails to translate into appropriate output behavior—revealing a disconnect between representation and decision-making in these models.

Method	Layer	Acc	F1	AUC
SVM	1	48.2	51.0	53.8
	16	93.4	93.6	97.9
	32	95.8	95.9	98.8
Random Forest	1	56.4	58.6	60.4
	16	97.4	97.4	99.5
	32	98.0	98.0	99.8

Table 3: **Subspace Classification Performance.** We train SVM and Random Forest Classifier on the adversarial and faithful input.

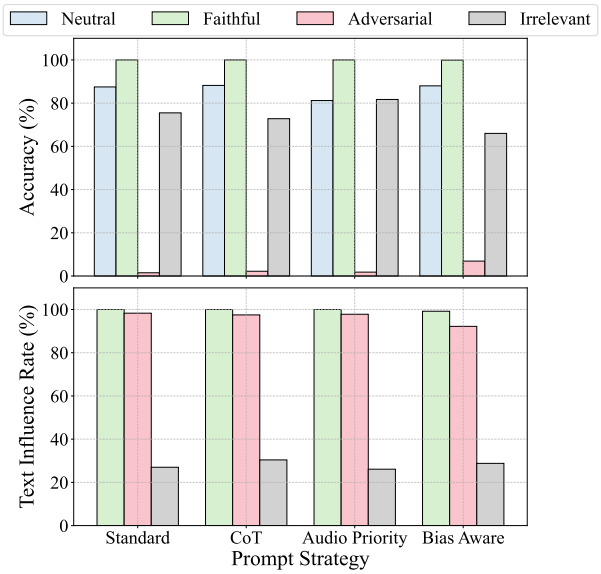

Figure 6: **Analysis on Different Prompting Techniques.** We perform our experiments on Qwen2-Audio-Instruct and MCR-BENCH subset.

6 Mitigating Text Bias

We discuss two potential solutions to mitigate the text bias in LALMs.

6.1 Prompting Techniques

Inspired by previous studies (Shi et al., 2023; Deng et al., 2025), we first investigate whether different prompting techniques will help models reduce the text bias. We consider the following techniques: Zero-Shot Chain-of-Thought prompting (Kojima et al., 2023), Audio Priority prompting which explicitly instructs the model to prioritize audio information, and Bias Awareness prompting which reminds the model about potential modality conflicts (prompts are shown in Appendix B).

In our experiments with Qwen2-Audio-Instruct on the Audio Question Answering subset of MCR-BENCH (result in Figure 6), we find that prompting techniques alleviate the text bias to some extent, but the improvement is very limited. Specifically, the Bias Awareness prompt shows the most signifi-

Method	ClothoAQA				MELD				VocalSound (Out-of-Distribution)			
	Acc_{fth}	Acc_{adv}	Acc_{irr}	TIR_{adv}	Acc_{fth}	Acc_{adv}	Acc_{irr}	TIR_{adv}	Acc_{fth}	Acc_{adv}	Acc_{irr}	TIR_{adv}
Base	100.0	1.5	75.5	98.3	100.0	0.0	27.8	100.0	99.8	11.8	85.7	86.2
Prompt-based Methods												
w/ CoT	100.0	2.2	72.8	97.5	100.0	0.0	28.5	100.0	99.8	11.9	84.9	86.1
w/ Bias Awareness	99.9	6.9	66.0	92.2	100.0	0.0	28.0	100.0	100.0	11.8	85.1	86.3
w/ Audio Priority	100.0	1.8	81.7	97.8	100.0	0.0	26.2	100.0	99.9	12.1	85.2	86.0
Best Prompt	100.0	6.9	81.7	92.2	100.0	0.0	28.5	100.0	10.0	12.1	85.7	86.3
Finetuning-based Methods												
w/ SFT	90.9	42.1	89.2	18.7	60.6	43.8	47.2	14.6	96.2	17.7	92.1	76.9

Table 4: **Comparison of techniques for mitigating text bias in Qwen2-Audio-Instruct.** We compare the base model, bias awareness prompting, and SFT across training datasets and evaluate generalization on the out-of-distribution subset.

cant effect, increasing the accuracy from 1.5% to 17.4% under adversarial conditions. It also decreases the Text Influence Rate from 98.3% to 79.7%, indicating reduced susceptibility to misleading text. However, even with these improvements, the model's performance remains compromised when faced with contradictory textual information, suggesting that more fundamental architectural or training modifications may be necessary to effectively address the text bias problem in LALMs.

6.2 Supervised Finetuning (SFT)

We investigate whether supervised finetuning (SFT) on datasets containing conflicting audio-text pairs can mitigate text bias in LALMs. This strategy explicitly trains the model to recognize and resolve cross-modal inconsistencies by providing the correct answers despite misleading textual information. This targeted intervention aims to recalibrate the model's attention between modalities when faced with conflicting inputs.

We use Qwen2-Audio-Instruct as our base model, and fine-tune it on 1,000 samples from Audio Question Answering and Speech Emotion Recognition subsets that contain deliberately mismatched audio-text pairs. To ensure efficient adaptation while preserving general capabilities, we employ Low-Rank Adaptation (LoRA) (Hu et al., 2022) with a rank of 8 and train for 2 epochs. Finetuning details are given in Appendix C. We evaluate the model's generalization on the Vocal Sound Classification task, which represents an unseen domain.

Table 4 presents the performance of the base and fine-tuned models across different metrics. We observe that SFT substantially outperforms prompt-based methods in mitigating text bias. Our fine-tuned model shows dramatically improved adversarial accuracy across all datasets, with particularly notable gains on Audio Question Answering and Speech Emotion Recognition tasks. This comes with a significant reduction in Text Influence Rate, indicating enhanced resistance to misleading textual cues. However, this improvement trades off some performance on faithful text conditions, suggesting a recalibration of modality attention rather than an overall enhancement. Interestingly, the model exhibits improved handling of irrelevant textual inputs as well, demonstrating more balanced cross-modal processing. Despite these gains, text bias remains present, highlighting the need for more advanced architectural approaches to fully resolve modality imbalance in LALMs.

7 Conclusion

In this work, we introduce MCR-BENCH, a benchmark that evaluates the performance of LALMs when faced with cross-modal inconsistencies. Our comprehensive evaluations across multiple models and tasks demonstrate that state-of-the-art LALMs exhibit a strong bias towards textual input over audio, leading to consistent performance degradation under adversarial conditions. We explore various mitigation strategies, which can only partially address the issue. These findings highlight the critical reliability concerns for real-world applications and underscore the need for novel training paradigms to better balance modality contributions in multimodal processing. We believe MCR-BENCH will serve as a valuable benchmark for developing more robust large audio-language models.

Limitations

Despite our comprehensive evaluation, this study has several limitations. Our analysis is constrained to specific audio understanding tasks and may not generalize to all audio-language scenarios. The synthetic nature of our adversarial and irrelevant textual descriptions might present different challenges compared to naturally occurring conflicts. Our investigation of mitigation strategies was limited to prompting techniques and model scaling, without exploring architectural modifications or specialized training objectives that could potentially yield more substantial improvements. Additionally, our evaluation focused on English-language models and Western audio contexts, potentially missing cultural and linguistic factors that may influence cross-modal processing priorities.

Acknowledgments

This research is supported by the National Research Foundation, Singapore and Infocomm Media Development Authority under its Trust Tech Funding Initiative. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore and Infocomm Media Development Authority. This work is supported by the National Research Foundation, Singapore, and Cyber Security Agency of Singapore under its National Cybersecurity R&D Programme and CyberSG R&D Cyber Research Programme Office. Any opinions, findings and conclusions or recommendations expressed in these materials are those of the author(s) and do not reflect the views of National Research Foundation, Singapore, Cyber Security Agency of Singapore as well as CyberSG R&D Programme Office, Singapore.

References

Tincans AI. 2024. Gazelle: Joint speech-language model. https://github.com/tincans-ai/gazelle. Version 0.2.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *arXiv preprint arXiv:2006.11477*.

Nicholas Carlini and David Wagner. 2018. Audio adversarial examples: Targeted attacks on speech-to-text. In *2018 IEEE Security and Privacy Workshops (SPW)*.

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

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

Ailin Deng, Tri Cao, Zhirui Chen, and Bryan Hooi. 2025. Words or vision: Do vision-language models have blind faith in text? *arXiv preprint arXiv:2503.02199*.

Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. 2019. Clotho: An audio captioning dataset.

Sreyan Ghosh, Zhifeng Kong, Sonal Kumar, S Sakshi, Jaehyeon Kim, Wei Ping, Rafael Valle, Dinesh Manocha, and Bryan Catanzaro. 2025. Audio flamingo 2: An audio-language model with long-audio understanding and expert reasoning abilities.

Yuan Gong, Alexander H. Liu, Hongyin Luo, Leonid Karlinsky, and James Glass. 2023. Joint audio and speech understanding.

Yuan Gong, Jin Yu, and James Glass. 2022. Vocalsound: A dataset for improving human vocal sounds recognition. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 151–155.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Yue Huang, Yanbo Wang, Zixiang Xu, Chujie Gao, Siyuan Wu, Jiayi Ye, Xiuying Chen, Pin-Yu Chen, and Xiangliang Zhang. 2025. Breaking focus: Contextual distraction curse in large language models.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. 2019. AudioCaps: Generating captions for audios in the wild. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 119–132, Minneapolis, Minnesota. Association for Computational Linguistics.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2023. Large language models are zero-shot reasoners.

Samuel Lipping, Parthasaarathy Sudarsanam, Konstantinos Drossos, and Tuomas Virtanen. 2022. Clotho-aqa: A crowdsourced dataset for audio question answering.

Chaoqun Liu, Mahani Aljunied, Guizhen Chen, Hou Pong Chan, Weiwen Xu, Yu Rong, and Wenxuan Zhang. 2025a. Seallms-audio: Large audio-language

models for southeast asia. https://github.com/DAMO-NLP-SG/SeaLLMs-Audio.

Ming Liu, Hao Chen, Jindong Wang, and Wensheng Zhang. 2025b. On the robustness of multimodal language model towards distractions. *arXiv preprint arXiv:2502.09818*.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Annamaria Mesaros, Toni Heittola, Tuomas Virtanen, and Mark D. Plumbley. 2021. Sound event detection: A tutorial. *IEEE Signal Processing Magazine*, 38(5):67–83.

OpenAI. 2024. Gpt-4o system card.

Daniel S Park, William Chan, Yu Zhang, et al. 2019. Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779*.

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

Yu Qin, Nicholas Carlini, Ekin D Cubuk, et al. 2019. Imperceptible, robust, and targeted adversarial examples for automatic speech recognition. *arXiv preprint arXiv:1902.08790*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning*, pages 31210–31227. PMLR.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. 2023. Salmonn: Towards generic hearing abilities for large language models. *arXiv preprint arXiv:2310.13289*.

Bin Wang, Xunlong Zou, Geyu Lin, Shuo Sun, Zhuohan Liu, Wenyu Zhang, Zhengyuan Liu, AiTi Aw, and Nancy F Chen. 2025. Audiobench: A universal benchmark for audio large language models. *NAACL*.

Zijun Wu, Bingyuan Liu, Ran Yan, Lei Chen, and Thomas Delteil. 2024. Reducing distraction in long-context language models by focused learning.

Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, Chang Zhou, and Jingren Zhou. 2024. AIR-bench: Benchmarking large audio-language models via generative comprehension. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1979–1998, Bangkok, Thailand. Association for Computational Linguistics.

A Faithful and Adversarial Statement Generation

The prompt used to generate faithful statements that accurately reflect audio content and adversarial statements that contradict the audio content is presented in Figure 7.

Prompt for Text Variants Generation

Convert this question and answer into two statements:
1. A factual statement that accurately represents the information from the question and answer.
2. A non-factual statement that contradicts the factual statement.

Example:
Question: "Are people speaking?"
Answer: "yes"
Factual statement: "There are people speaking."
Non-factual statement: "There are no people speaking."

Now convert this pair:
Question: "<QUESTION>"
Answer: "<ANSWER>"
Factual statement:
Non-factual statement:

Figure 7: Prompt used for generating text variants from question-answer pairs.

B Mitigation Strategy Prompts

This section details the prompt used for mitigating text bias of LALMs (shown in Figure 8 and Figure 9).

Audio Priority Prompt

After thinking step by step, please answer the question.

Figure 8: CoT prompt for mitigating text bias.

Audio Priority Prompt

Please prioritize information from the audio over the text description.

Figure 9: Audio Priority prompt for mitigating text bias.

Bias Awareness Prompt

Be aware that you may have a tendency to trust text descriptions more than audio evidence. Try to avoid this text bias and then answer the question.

Figure 10: Bias Awareness prompt for mitigating text bias.

C SFT Details

We fine-tuned the Qwen2-Audio-7B-Instruct model using LoRA with rank 8 and $\alpha = 32$, targeting all linear layers while freezing the ViT components. Training ran for 2 epochs with a learning rate of 1e-4 and warmup ratio of 0.05. We used a per-device batch size of 1 with gradient accumulation steps of 16, resulting in an effective batch size of 128. All training was performed using bfloat16 precision with a maximum sequence length of 2048 tokens.