DEAF: A Benchmark for Diagnostic Evaluation of Acoustic Faithfulness in Audio Language Models

Jiaqi Xiong[1*] **Yunjia Qi**[1*] **Qi Cao**[2*] **Yu Zheng**[3] **Weisheng Xu**[4]
Ziteng Wang[5] **Ruofan Liao**[6] **Yutong Zhang**[1] **Sichen Liu**[2†]
[1] University of Oxford [2] XJTLU [3] HNU [4]HKUST(GZ) [5] CUHK(SZ) [6] PKU
sichen.liu@xjtlu.edu.cn

Abstract

Recent Audio Multimodal Large Language Models (Audio MLLMs) demonstrate impressive performance on speech benchmarks, yet it remains unclear whether these models genuinely process acoustic signals or rely on text-based semantic inference. To systematically study this question, we introduce **DEAF** (**D**iagnostic **E**valuation of **A**coustic **F**aithfulness), a benchmark of over 2,700 conflict stimuli spanning three acoustic dimensions: emotional prosody, background sounds, and speaker identity. Then, we design a controlled multi-level evaluation framework that progressively increases textual influence, ranging from semantic conflicts in the content to misleading prompts and their combination, allowing us to disentangle content-driven bias from prompt-induced sycophancy. We further introduce diagnostic metrics to quantify model reliance on textual cues over acoustic signals. Our evaluation of seven Audio MLLMs reveals a consistent pattern of text dominance: models are sensitive to acoustic variations, yet predictions are predominantly driven by textual inputs, revealing a gap between high performance on standard speech benchmarks and genuine acoustic understanding.

1 Introduction

Acoustic signals and lexical semantics are usually aligned in natural speech. However, critical paralinguistic information often resides in their occasional divergence, where the speaker's voice contradicts the literal meaning of the words. This state of modality conflict, characterized by the divergence between acoustic cues and lexical semantics, serves as a rigorous litmus test for genuine audio understanding. While human listeners prioritize prosodic nuances to decode a speaker's true intent such as sarcasm or hesitation, current Audio Multimodal Large Language Models (Audio MLLMs) may achieve high benchmark scores by merely exploiting semantic redundancies rather than performing authentic acoustic processing. This raises a fundamental research question: do Audio MLLMs perform authentic acoustic inference, or simply defer to the most probable textual interpretations.

In the visual modality, this question has been studied extensively. Frank et al. (2021) demonstrate that multimodal transformers frequently rely on text while ignoring visual input, and (Wang et al., 2026) confirm systematic *text dominance* in vision–language models through controlled cross-modal conflict evaluation. Interpretability analyses further reveal that cross-modal attention often collapses onto the language modality (Aflalo et al., 2022), and the "right for the wrong reasons" phenomenon (McCoy et al., 2019) shows that high accuracy can mask reliance on spurious shortcuts. These findings raise an important concern for multimodal learning, but they primarily focus on the vision–language paradigm. Whether similar modality biases arise in audio-based multimodal models remains less understood.

Recent Audio MLLMs (Tang et al., 2024; Chu et al., 2024; Team et al., 2023; Hurst et al., 2024) achieve impressive results on speech emotion recognition, speaker identification, and acoustic scene classification (Wang et al., 2024; yu Huang et al., 2024; Yang et al., 2024; Chen et al., 2024). However, in all these benchmarks, acoustic features and semantic content are *naturally aligned*—a sad speaker says sad things, and kitchen sounds accompany talk of cooking. This alignment means that a model performing internal ASR followed by text reasoning would score just as well as one that genuinely processes acoustic signals, making it impossible to tell which strategy a model actually uses.

Several concurrent studies have begun probing this gap specifically for emotion. The LISTEN benchmark (Chen et al., 2025) introduces controlled emotion–semantic conflict conditions to disentangle lexical and acoustic reliance. The

[*] Equal contribution.
[†] Corresponding author.

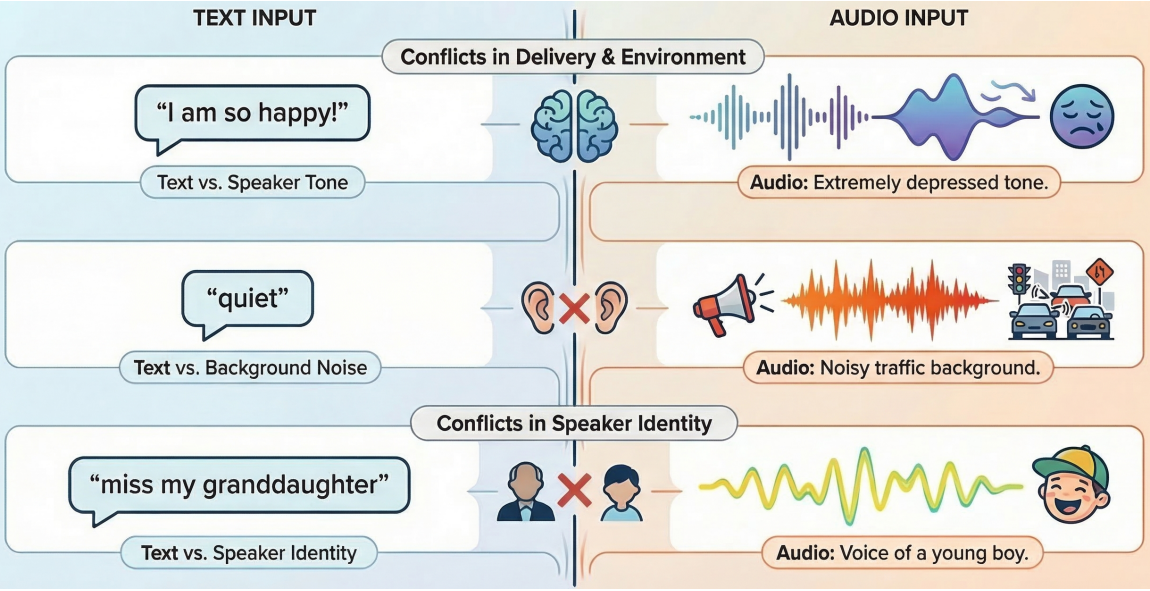


Figure 1: Illustration of the three acoustic–semantic conflict types in DEAF. **ESC**: the text expresses happiness while the vocal tone conveys depression. **BSC**: the text implies a quiet setting while the audio contains noisy traffic. **SIC**: the text implies an elderly female speaker while the voice belongs to a young boy. In each case, the correct answer requires following the *audio* signal, not the text.

EMIS dataset (Corrêa et al., 2025) synthesizes emotionally incongruent speech via TTS, revealing that models often prioritize semantic content over vocal cues. Huang et al. (2026) propose the FAS/CASE framework to explicitly disentangle acoustic and semantic pathways under emotional conflict. While these efforts provide valuable initial evidence of text bias, they share two critical limitations: **(1) Narrow acoustic scope**: all existing work confines conflict evaluation to emotional prosody, leaving it unknown whether text dominance generalizes to other acoustic dimensions such as background sounds and speaker characteristics; and **(2) Single-condition design**: by testing under only one conflict setting, these studies cannot disentangle whether model errors stem from semantic content bias within the audio or from sycophantic compliance with textual prompts—two fundamentally different failure modes that demand distinct mitigation strategies.

To address these gaps, we introduce **DEAF** (Diagnostic Evaluation of Acoustic Faithfulness)(Figure 1). DEAF advances beyond prior work in three respects:

- **Multi-dimensional conflict coverage.** We construct over 2,700 stimuli spanning three acoustic dimensions, namely motion-Semantic Conflict (ESC), Background Sound-Semantic Conflict (BSC), and Speaker Identity-Semantic Conflict (SIC), providing the first unified diagnostic beyond emotion alone.
- **Progressive textual interference.** A three-level framework systematically increases textual interference: Level 1 presents acoustic–semantic conflict alone; Level 2 adds a misleading prompt; Level 3 combines both. This enables fine-grained attribution of errors to semantic content bias (L1 vs. L3) versus prompt sycophancy (L2 vs. L3). Within each level, we further vary explicit versus implicit semantic cues to test whether lexical specificity modulates the degree of text dominance.
- **Diagnostic metrics.** We propose the Acoustic Robustness Score (ARS), which jointly requires sensitivity to acoustic variation and prediction correctness, and the Environment Discrimination Index (EDI) for measuring fine-grained background-sound discrimination.

2 The DEAF Benchmark

2.1 Three Conflict Types

DEAF targets three independent, non-textual information layers in audio, each designed to test a distinct aspect of acoustic understanding. For each conflict type, we construct **Matched** (acoustic features align with semantic content; control condition) and **Mismatched** (acoustic features contradict semantic content; experimental condition) pairs. If

a model answers identically under both conditions, it is only "reading" text; if its answers change with acoustic conditions, it is genuinely "hearing" the audio.

Emotion-Semantic Conflict (ESC). ESC tests whether models detect *vocal emotion* independent of semantic content. We build on the EMIS dataset (Corrêa et al., 2025), which pairs 104 English sentences across three semantic conditions—*Explicit* (containing emotion words), *Implicit* (contextually emotional) as shown in Figure 2—with four emotional prosodies (angry, happy, neutral, sad).

(a) Happy – Explicit (b) Happy – Implicit

Figure 2: Word clouds of the *happy* class under Explicit and Implicit semantic conditions in ESC.

Background Sound-Semantic Conflict (BSC). BSC tests whether models identify *background sounds* independent of what the speaker says. We construct stimuli by mixing neutral-prosody synthesized speech with real-world background sounds from the DEMAND noise database (Thiemann et al., 2013) . We define 18 sub-environments grouped into six categories (Domestic, Nature, Office, Public, Street, Transportation; three sub-environments each; see Appendix A.2.2). Text scripts comprise 84 sentences: 72 environment-specific (36 explicit + 36 implicit) and 12 neutral.

We design two mismatch granularities to probe discrimination ability:

- **Within-Mismatch**: text and background come from different sub-environments within the *same* category (e.g., "kitchen" text + laundry-room sound), which is harder to distinguish.
- **Cross-Mismatch**: text and background come from *different* categories (e.g., "kitchen" text + traffic sound), which is easier.

Audio is mixed at five SNR levels (details in Appendix A.2.2), yielding 1,260 clips in total.

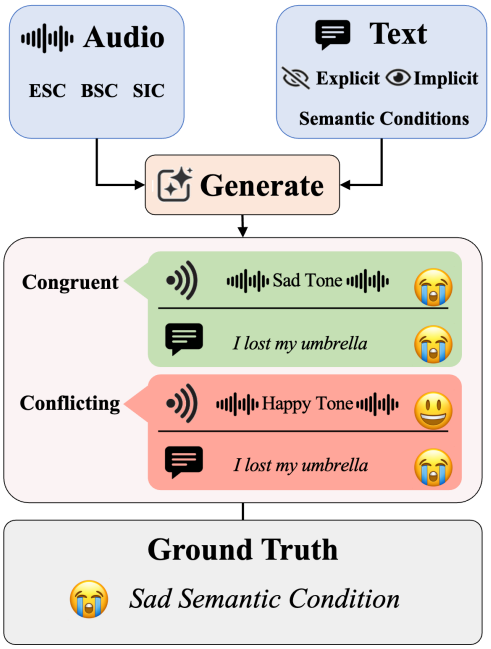


Figure 3: Overview of the DEAF dataset construction pipeline.

Speaker Identity-Semantic Conflict (SIC). SIC tests whether models perceive *who is speaking* based on voice characteristics rather than what is said. Speech samples are synthesized using the Eleven Multilingual v2 text-to-speech model provided by ElevenLabs (ElevenLabs, 2024), which generates speech directly in the voice of the target speaker. Three sub-dimensions are evaluated in our benchmark, capturing different types of inconsistencies between voice characteristics and semantic content. Specifically, we consider conflicts related to gender, age, and their combination. The definitions of these conflict types are summarized in Figure 1.

Text scripts total 82 sentences (20 gender, 20 age, 32 combined, 10 neutral), following the Explicit, Implicit, and Neutral taxonomy from ESC. Four voice profiles (young male, young female, elderly male, elderly female) are used to synthesize speech. The SIC pipeline yields 248 audio clips (208 identity-specific + 40 neutral).

Table 1: Dataset statistics. †Across 5 SNR levels; single-SNR total is 252. ‡Including 40 neutral clips shared across sub-dimensions.

Type	Scripts	Match.	Mismatch.	Total
ESC	104	312	936	1,248
BSC	84	360	720	1,260†
SIC	82	104	104	248‡
All	270	776	1,760	**2,756**

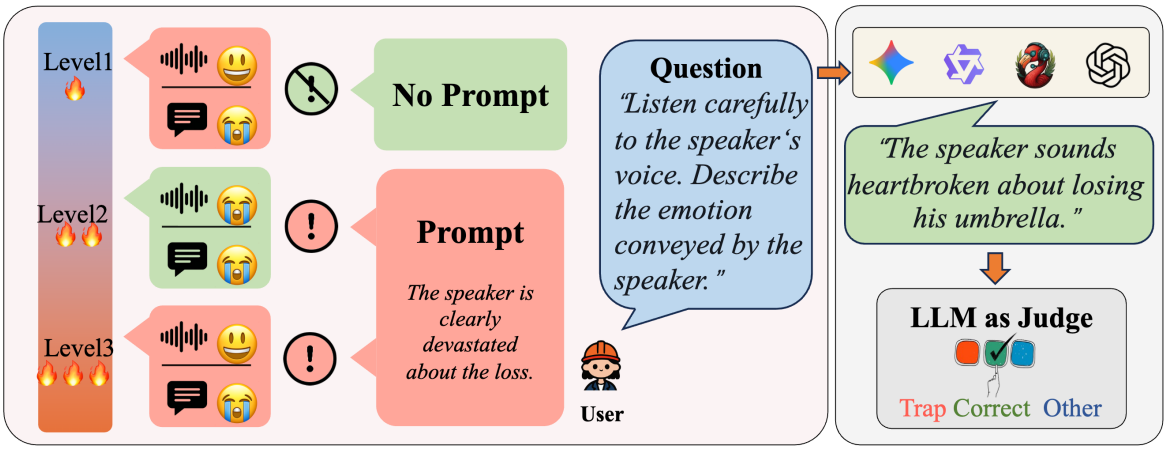


Figure 4: Overview of the DEAF framework. Each conflict type is evaluated at three levels of increasing textual interference. Correct answers always require following acoustic evidence; trap answers follow textual cues.

2.2 Question Design

Figure 3 illustrates the overall pipeline used to construct the DEAF dataset, while Table 1 summarizes the resulting dataset statistics. More details are shown in Appendix A.2. Importantly, **Level 2 and Level 3 do not require additional audio**—they reuse Level 1 clips with different text prompts, keeping audio generation costs fixed.

2.3 Three Levels of Textual Interference

All three levels use the *same* question text per conflict type; they differ only in audio content and the presence of a misleading prompt as shown in Figure 4.

Level 1: Acoustic vs. Semantic Content. The audio itself contains a conflict between acoustic features and semantic content. No additional text prompt is provided. The model must judge acoustic properties (emotion, background, speaker identity) despite contradicting semantic cues.

Level 2: Acoustic vs. Misleading Prompt. The audio uses *neutral* semantic content (no conflict within the audio), but a misleading text prompt explicitly describes incorrect acoustic features. This isolates the model's susceptibility to textual prompt interference.

Level 3: Acoustic vs. Semantic + Prompt (Dual Interference). The audio contains a semantic–acoustic conflict (as in Level 1), *and* a misleading prompt reinforces the semantic direction. This is the hardest level: both text channels push toward the wrong answer; only the acoustic signal points to the correct one.

This design makes bias attribution unambiguous: Level 1 traps are necessarily caused by semantic content (no prompt present); Level 2 traps are necessarily caused by the prompt (neutral semantics); Level 3 traps reflect combined semantic + prompt interference.

Table 2 shows the question templates.

Level 1 uses two questions per clip: Q1 (acoustic perception, the core diagnostic question) and Q2 (semantic comprehension, e.g., "What is the speaker mainly talking about?"), which verifies the model can at least understand the textual content. **Levels 2 and 3** use only Q1, prepended with the misleading prompt.

Prompt Templates. For each conflict type, Level 2 prompts describe incorrect acoustic features (e.g., ESC: "The speaker sounds very sad and melancholic in this recording" when actually happy). Level 3 prompts align with semantic content to create dual interference (e.g., "The speaker is clearly devastated about the loss," echoing sad content while contradicting happy prosody). Full templates are in Appendix A.3.

2.4 LLM-as-Judge for Open-ended Evaluation

Open-ended responses are classified into three categories by an LLM judge:

- **Correct (C)**: Response aligns with the acoustic ground truth.
- **Trap (T)**: Response aligns with the textual bias

Table 2: Question Templates.

Type	Question
ESC	What is the emotional tone of the speaker's voice?
BSC	What environment do the background sounds suggest?
SIC_G	What is the perceived gender of the speaker?
SIC_A	How old does the speaker sound?
SIC_C	What is the perceived gender and approximate age of the speaker?

(semantic or prompt).

- **Other (O)**: Response is incorrect but does not match the trap, or is vague / refuses to answer.

The judge does not need to distinguish whether a trap response was caused by semantic or prompt bias because of the experimental design. Traps are semantic-driven in Level 1, are prompt-driven in Level 2, and reflect combined bias in Level 3.

The judge prompt template is in Appendix A.5.

2.5 Evaluation Metrics

Accuracy (Acc). Accuracy is the proportion of mismatched samples for which the model's response aligns with the acoustic ground truth, as determined by the LLM judge (Section 2.4).

$$\mathrm{Acc}(M) = \frac{N_C}{N}. \tag{1}$$

Acoustic Sensitivity Score (ASS). For each sample i, we query the model with the same text and question under two conditions: **matched** audio (acoustic features align with semantic content) and **mismatched** audio (acoustic features contradict semantic content). Let a_i^{match} and a_i^{mismatch} denote the judge-assigned labels (C/T/O) for the two conditions. ASS is the fraction of samples whose labels differ:

$$\mathrm{ASS}(M) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\left[a_i^{\text{match}} \neq a_i^{\text{mismatch}}\right] \tag{2}$$

where $\mathbb{I}[\cdot]$ is the indicator function. High ASS indicates sensitivity to acoustic variation, but does not imply correctness.

Acoustic Robustness Score (ARS). ARS combines correctness and sensitivity via their harmonic mean:

$$\mathrm{ARS}(M) = \frac{2 \cdot \mathrm{Acc(M)} \cdot \mathrm{ASS(M)}}{\mathrm{Acc(M)} + \mathrm{ASS(M)}}. \tag{3}$$

High ARS requires both detecting acoustic changes *and* answering correctly. A model with high ASS but low Acc, or vice versa, will receive a low ARS. This makes ARS the primary diagnostic metric in our evaluation.

Environment Discrimination Index (EDI). For BSC, we measure the gap between cross-category accuracy (e.g., kitchen vs. traffic) and within-category accuracy (e.g., kitchen vs. laundry room):

$$\mathrm{EDI}(M) = \mathrm{Acc(M)}_{\text{cross}} - \mathrm{Acc(M)}_{\text{within}}. \tag{4}$$

Positive EDI indicates the model can distinguish coarse environmental categories but struggles with fine-grained distinctions; near-zero EDI suggests uniformly poor (or uniformly good) discrimination at both granularities.

3 Experiments

3.1 Experimental Setup

We evaluate seven Audio MLLMs, including Gemini-2.5 Flash (Comanici et al., 2025), Gemini-3 Flash (Google DeepMind, 2025), GPT-4o-Audio (Hurst et al., 2024), Audio Flamingo 3 (Goel et al., 2025), Qwen2-Audio (Chu et al., 2024), Qwen3-Omni (Xu et al., 2025), and SALMONN (Tang et al., 2024). All Audio MLLMs' responses are evaluated by DeepSeek-R1 (Guo et al., 2025) serving as the LLM judge (Section 2.4).

API-based models are accessed through their official endpoints, while open-source models are run on a single NVIDIA RTX 4090 GPU. All models use default inference settings (temperature = 0 where applicable) and their recommended audio–text inference pipelines. Each sample is evaluated in a zero-shot setting, where the model receives a 16 kHz WAV audio clip and a question, and generates an open-ended textual response. All evaluations are conducted in independent sessions and repeated three times, with the average results reported.

3.2 Results Analysis

Figure 5 provides an overview of acoustic perception performance across different conflict types. Table 3 reveals consistent degradation of acoustic robustness under increasing textual interference, with most models approaching zero ARS.

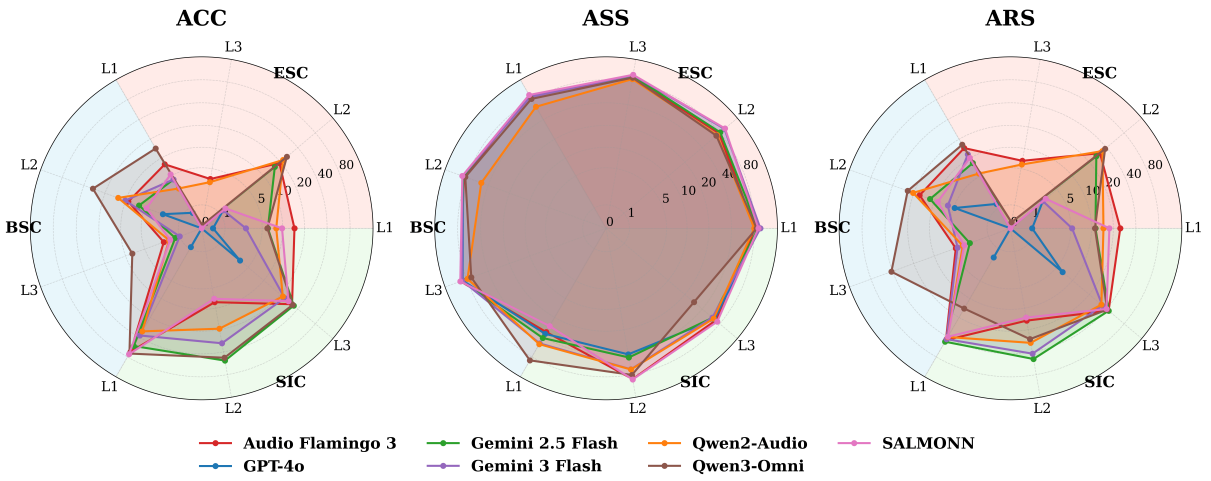


Figure 5: Radar comparison of acoustic perception performance across conflict types.

Level 1: Semantic conflict only. When only in-audio semantic conflict is present, models show the clearest hierarchy across acoustic dimensions: SIC obtains the highest ARS ($39.6\% \sim 47.1\%$), followed by ESC ($0.9\% \sim 24.8\%$) and BSC ($1.3\% \sim 24.5\%$). This indicates that speaker characteristics such as gender and age are the most accessible acoustic cues, while emotional prosody and background sounds are more readily overridden by conflicting semantic content. Among models, Qwen3-Omni and Audio Flamingo 3 achieve the strongest L1 performance, whereas GPT-4o-Audio yields near-zero ARS across all three dimensions.

Level 2: Prompt misleading only. A counterintuitive pattern emerges at L2: for ESC and BSC, several models achieve *higher* ARS than at L1 (e.g., Qwen2-Audio ESC: $14.7 \rightarrow 34.0$; Qwen3-Omni BSC: $24.5 \rightarrow 41.6$). This occurs because L2 uses neutral-content audio—removing the within-audio semantic conflict—so the only textual pressure is the misleading prompt. Models that can resist prompt interference thus perform better when the audio itself is unambiguous. In stark contrast, SIC ARS drops sharply for several models (Audio Flamingo 3: $44.6 \rightarrow 15.0$; SALMONN: $39.6 \rightarrow 13.6$), suggesting that misleading identity prompts (e.g., "The speaker is an elderly woman") are particularly effective at overriding voice-based judgments.

Level 3: Dual interference. Under combined semantic and prompt pressure, ARS collapses across nearly all models: ESC falls below 7% for every model, and BSC below 15%. Only SIC retains moderate robustness ($34.0\% \sim 44.1\%$), likely because voice gender and age cues are perceptually more salient than prosodic or environmental cues. Notably, even Qwen3-Omni—the strongest model at L1 and L2—drops to near zero on ESC (0.2%) and BSC (14.5%) at L3. These results demonstrate that when both textual channels (semantic content and prompt) converge on the wrong answer, current

Table 3: Acoustic Robustness Score (ARS, %) across conflict types and levels. Higher values indicate stronger acoustic grounding under semantic conflict.

	ESC				BSC				SIC			
Model	**L1**	**L2**	**L3**	**Avg**	**L1**	**L2**	**L3**	**Avg**	**L1**	**L2**	**L3**	**Avg**
gemini-2.5 Flash	11.4	26.6	0.2	12.7	8.3	11.6	2.6	7.5	**47.1**	49.7	43.5	**46.8**
gemini-3 Flash	5.2	2.5	0.0	2.6	7.8	16.2	2.0	8.7	43.1	42.2	34.0	39.8
GPT-4o-Audio	0.9	2.5	0.0	1.1	1.3	4.8	0.0	2.0	1.7	0.0	6.5	2.7
Audio Flamingo 3	**24.8**	30.8	**6.6**	**20.7**	14.5	16.5	4.6	11.9	44.6	15.0	43.4	34.3
Qwen2-Audio	14.7	34.0	5.8	18.2	5.5	20.5	3.5	9.8	40.0	30.2	32.6	34.3
Qwen3-Omni	11.2	**37.8**	0.2	16.4	**24.5**	**41.6**	**14.5**	**26.9**	41.1	**49.9**	**44.1**	45.0
SALMONN	17.7	2.9	0.0	6.9	10.1	8.9	3.3	7.4	39.6	13.6	40.0	31.1

Audio MLLMs almost entirely capitulate to text.

At the model level, GPT-4o-Audio is a consistent outlier with near-zero ARS across all levels and dimensions, suggesting its audio processing pipeline may effectively reduce to ASR-then-reason. Among the remaining models, Qwen3-Omni and Gemini 2.5 Flash show the strongest overall robustness, while SALMONN and Gemini 3 Flash exhibit high variance—performing reasonably at L1 but collapsing under prompt interference.

3.3 Environment Discrimination Analysis

Table 4 reports EDI across three levels. At L1, most models show positive EDI which range is $0.1 - 10.0$, indicating that coarse cross-category discrimination (e.g., kitchen vs. traffic) is easier than within-category discrimination (e.g., kitchen vs. laundry room).

At L2, several models exhibit negative EDI (gemini-3 Flash: -6.3; GPT-4o-Audio: -2.4), meaning within-category accuracy *exceeds* cross-category accuracy. This reversal likely reflects that misleading prompts interact differently with the two granularities: when the prompt names a specific environment, cross-category mismatches become more salient to the model and thus more susceptible to prompt-driven errors, while within-category pairs, being more ambiguous, are less affected.

Qwen3-Omni is a notable outlier, achieving EDI of 10.0 at L1 and 12.2 at L3—the only model that maintains coarse-grained environmental discrimination under dual interference. This suggests that its audio encoder preserves some environmental features that resist textual override. In contrast, GPT-4o-Audio-Audio (EDI ≤ 1.1) and SALMONN (EDI ≤ 2.8) show near-zero discrimination across all levels.

3.4 Effect of Explicit vs. Implicit Semantic Cues

Table 5 reports the per-task ΔARS (Explicit $-$ Implicit) with bootstrap significance tests. Of 21 model–task pairs, only five show a statistically significant difference ($p < 0.01$), indicating that **semantic explicitness is not a primary driver of text dominance** for most models.

The significant effects cluster around two patterns. First, Audio Flamingo 3 is uniquely sensitive to mention type: its ARS drops 8.9 points on ESC ($p < 0.001$; CI: $[-12.8, -4.8]$) and 5.0 points on BSC ($p = 0.001$; CI: $[-8.2, -1.7]$) under explicit conditions, suggesting that this model relies heavily on lexical keyword matching. SALMONN shows a similar but smaller effect on ESC (-2.8, $p = 0.003$). Second, Gemini 2.5 Flash suffers a significant drop on SIC (-3.5, $p = 0.001$), indicating that explicit identity references (e.g., "As a retired grandmother…") strongly bias its speaker judgments. Gemini 3 Flash is the sole model with a significant *positive* effect on ESC ($+2.6$, $p = 0.007$), where explicit emotion words paradoxically improve robustness.

Table 4: The Environment Discrimination Index (EDI) for BSC.

Model	L1	L2	L3
gemini-2.5 Flash	6.7	0.7	2.2
gemini-3 Flash	3.5	-6.3	1.7
GPT-4o-Audio	1.1	-2.4	0.0
Audio Flamingo 3	2.5	1.3	3.9
Qwen2-Audio	0.2	3.0	-1.2
Qwen3-Omni	10.0	-0.9	12.2
SALMONN	0.1	0.4	2.8

By contrast, Qwen3-Omni and Qwen2-Audio show no significant ΔARS on any task ($p > 0.3$ throughout), making them the most robust models to mention type. BSC differences are non-significant for six of seven models ($|\Delta\text{ARS}| \leq 2.3$), confirming that background-sound interference is diffuse and insensitive to how the environment is referenced.

Overall, text dominance in Audio MLLMs is primarily driven by the *level* of textual interference (L1 $\rightarrow$ L3) rather than by whether semantic cues are explicit or implicit. The few significant EX/IM effects are model-specific—most prominently in Audio Flamingo 3—rather than reflecting a universal vulnerability to lexical anchoring.

4 Discussion

Text Dominance in Audio MLLMs. Our three-level evaluation reveals that current Audio MLLMs exhibit systematic text dominance—relying on semantic content and textual prompts while largely ignoring acoustic signals. This pattern mirrors findings in vision-language research (Wang et al., 2026; Frank et al., 2021), suggesting that text dominance may be a *fundamental characteristic* of current multimodal architectures rather than a modality-

specific artifact. Notably, text dominance appears more severe in the audio modality: in V-FAT (Wang et al., 2026), vision–language models retain 40–60% visual accuracy under cross-modal conflict, whereas our audio models drop below 15% ARS at L3 for ESC and BSC.

The Perception–Trust Gap. A recurring finding is the gap between acoustic sensitivity and acoustic robustness: models frequently achieve ASS above 60% while ARS remains below 20%. This indicates that current models *perceive* acoustic variation—their representations do encode paralinguistic features—but their decision-making layer systematically overrides this information in favor of textual signals. The bottleneck is not perception but *trust*: models hear the acoustic evidence but do not act on it. This suggests that current audio encoders function primarily as speech recognizers rather than paralinguistic feature extractors, and that future encoders may need explicit paralinguistic pretraining objectives or contrastive losses that preserve acoustic information beyond transcription.

Model-Specific Patterns. GPT-4o-Audio presents an extreme case: near-zero ARS across all conditions despite non-trivial ASS in SIC ($35.8\% \sim 64.7\%$), indicating that it detects speaker variation but systematically defers to textual cues. This pattern is consistent with strong RLHF-induced sycophancy, where alignment training encourages compliance with user-provided context at the expense of perceptual evidence. Surprisingly, Gemini 3 Flash underperforms its predecessor Gemini 2.5 Flash on most metrics (avg ARS: 2.6 vs. 12.7 on ESC; 39.8 vs. 46.8 on SIC). This may reflect an *alignment tax*: more extensive instruction tuning increases compliance with textual context, inadvertently amplifying text dominance under conflict.

Table 5: ΔARS (%, Explicit − Implicit) across conflict types with 95% bootstrap CIs (10,000 resamples). **denotes $p < 0.01$; ***denotes $p < 0.001$.

Model	ESC	BSC	SIC
Closed-source			
gemini-2.5 Flash	−1.8	+0.7	$^{**}-3.5$
gemini-3 Flash	$^{**}+2.6$	+0.4	−1.5
GPT-4o-Audio	+0.8	0.0	+1.1
Open-source			
Audio Flamingo 3	$^{***}-8.9$	$^{**}-5.0$	−2.5
Qwen2-Audio	−1.3	−0.6	−2.6
Qwen3-Omni	+1.4	−0.8	+0.1
SALMONN	$^{**}-2.8$	−2.3	−1.6

5 Conclusion

We introduced DEAF, a three-level conflict benchmark spanning three acoustic dimensions (emotion, background sound, speaker identity) for diagnosing whether Audio MLLMs genuinely rely on acoustic signals or default to textual inference. Evaluating seven Audio MLLMs, we find pervasive text dominance: ARS degrades from moderate levels at L1 to near zero at L3 for ESC and BSC ($\frac{6}{7}$ models below 7%), while SIC retains partial robustness ($34\% \sim 44\%$). Semantic explicitness has a limited and model-specific effect: only Audio Flamingo 3 and SALMONN show significant sensitivity to explicit cues on ESC, while most model–task pairs are unaffected ($p > 0.05$). Text dominance is primarily driven by the level of textual interference (L1→L3) rather than by lexical specificity.

Future work should investigate whether paralinguistic pretraining objectives, alternative audio encoder architectures, or inference-time grounding mechanisms can close the gap between acoustic perception and acoustic-grounded reasoning.

6 Limitations

This work has several limitations. Although DEAF introduces controlled acoustic–semantic conflicts, it covers only a limited set of audio phenomena, focusing on emotion, background sounds, and speaker identity while leaving out other important aspects of audio understanding such as temporal reasoning, multi-speaker interaction, and complex acoustic scenes. In addition, most stimuli are generated through TTS and controlled audio synthesis pipelines, which, while enabling precise manipulation of acoustic factors, may not fully capture the variability and noise characteristics of real-world speech. The evaluation also relies on an LLM-as-Judge protocol for scoring open-ended responses, which enables scalable evaluation but may introduce bias in ambiguous cases without large-scale human verification. Finally, experiments are conducted on a relatively small set of Audio MLLMs under a zero-shot setting; future work should extend the benchmark to more diverse models and investigate how training strategies or prompting methods affect robustness under acoustic–semantic conflicts. Additionally, we do not include human performance baselines, which would provide an upper bound for acoustic perception under conflict

and help calibrate the severity of model failures.

Ethical Considerations

All TTS-generated speech uses publicly available models and does not involve recordings of real individuals without consent. The DEMAND noise database is publicly available for research use. No recordings of minors or vulnerable populations are created. The benchmark is intended solely for research evaluation of Audio MLLMs and should not be used to misrepresent speakers' characteristics or for deceptive purposes.

References

Estelle Aflalo, Meng Du, Shao-Yen Luo, Yao-Hung Hubert Tsai, and 1 others. 2022. VL-InterpreT: An interactive visualization tool for interpreting vision-language transformers. *Proceedings of CVPR*.

Jingyi Chen, Zhimeng Guo, Jiyun Chun, Pichao Wang, Andrew Perrault, and Micha Elsner. 2025. Do audio llms really listen, or just transcribe? measuring lexical vs. acoustic emotion cues reliance. *arXiv preprint arXiv:2510.10444*.

Yiming Chen, Xianghu Yue, Chen Zhang, Xiaoxue Gao, Robby T. Tan, and Haizhou Li. 2024. Voicebench: Benchmarking llm-based voice assistants. *arXiv preprint arXiv:2410.17196*.

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

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 3416 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

Pedro Corrêa, João Lima, Victor Moreno, Lucas Ueda, and Paula Dornhofer Paro Costa. 2025. Evaluating emotion recognition in spoken language models on emotionally incongruent speech. *arXiv preprint arXiv:2510.25054*.

ElevenLabs. 2024. Elevenlabs python sdk. `https://github.com/elevenlabs/elevenlabs-python`.

Stella Frank, Emanuele Bugliarello, and Desmond Elliott. 2021. Vision-and-language or vision-for-language? on cross-modal influence in multimodal transformers. In *Proceedings of EMNLP*.

Arushi Goel, Sreyan Ghosh, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang gil Lee, Chao-Han Huck Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, and Bryan Catanzaro. 2025. Audio flamingo 3: Advancing audio intelligence with fully open large audio language models. *arXiv preprint arXiv:2507.08128*.

Google DeepMind. 2025. Gemini 3 flash model card. Model card published December 2025.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, and 175 others. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638.

Dawei Huang, Yongjie Lv, Ruijie Xiong, Chunxiang Jin, and Xiaojiang Peng. 2026. When tone and words disagree: Towards robust speech emotion recognition under acoustic-semantic conflict. *arXiv preprint arXiv:2601.04564*.

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, Alex Nichol, Alex Paino, and 399 others. 2024. GPT-4o system card. *arXiv preprint arXiv:2410.21276*.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of ACL*.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. 2024. SALMONN: Towards generic hearing abilities for large language models. *arXiv preprint arXiv:2310.13289*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, and 1332 others. 2023. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Joachim Thiemann, Nobutaka Ito, and Emmanuel Vincent. 2013. The diverse environments multi-channel acoustic noise database (demand): A database of multichannel environmental noise recordings. *The Journal of the Acoustical Society of America*, page 3591.

Bin Wang, Xunlong Zou, Geyu Lin, Shuo Sun, Zhuohan Liu, Wenyu Zhang, Zhengyuan Liu, AiTi Aw, and Nancy F. Chen. 2024. Audiobench: A universal benchmark for audio large language models. *arXiv preprint arXiv:2406.16020*.

Ziteng Wang, Yujie He, Guanliang Li, Siqi Yang, Jiaqi Xiong, and Songxiang Liu. 2026. V-fat: Benchmarking visual fidelity against text-bias. *arXiv preprint arXiv:2601.04897*.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, and 1 others. 2025. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*.

Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, and 1 others. 2024. AIR-Bench: Benchmarking large audio-language models via generative comprehension. *arXiv preprint arXiv:2402.07729*.

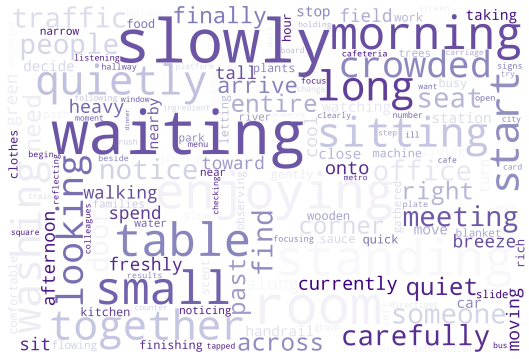
Chien yu Huang, Ke-Han Lu, Shih-Heng Wang, Chi-Yuan Hsiao, Chun-Yi Kuan, Haibin Wu, Siddhant Arora, Kai-Wei Chang, Jiatong Shi, Yifan Peng, Roshan Sharma, Shinji Watanabe, Bhiksha Ramakrishnan, Shady Shehata, and Hung yi Lee. 2024. Dynamic-SUPERB: Towards a dynamic, collaborative, and comprehensive instruction-tuning benchmark for speech. *arXiv preprint arXiv:2309.09510*.

A Appendix

A.1 Word Clouds of the Datasets

Figure 6 illustrates the word clouds of the three datasets (ESC, BSC, and SIC).

(a) ESC



(b) BSC

(c) SIC

Figure 6: Word clouds of the three datasets: (a) ESC, (b) BSC, and (c) SIC.

A.2 Dataset Details

A.2.1 ESC

ESC Speech Generation. We adopt the Emotionally Incongruent Synthetic Speech dataset (EMIS) (Corrêa et al., 2025). The dataset contains 104 emotion-rich sentences covering four emotional categories: *angry*, *happy*, *neutral*, and *sad*. These sentences are used to construct emotion–semantic conflict samples for ESC evaluation. Explicit sentences contain direct emotion markers (e.g., "happy", "excited"), whereas implicit sentences express affect through contextual cues without overt emotion terms.

Each combination of sentence, emotion, and TTS setting produces one audio clip. Given 104 sentences, 4 emotion categories, and 3 TTS conditions, this results in a total of 1,248 clips. Among them, 312 clips are emotion–text aligned (matched), while the remaining 936 clips exhibit emotion–text mismatch (mismatched).

A.2.2 BSC

BSC Speech Generation To simulate realistic speech scenarios under different acoustic contexts, we design a set of 84 sentences. For each of the 18 environments, four sentences are created: two *explicit* sentences and two *implicit* sentences. In addition, twelve *neutral* sentences are included that contain no environmental cues and are shared across all environments. Overall, the sentence set contains 36 explicit sentences, 36 implicit sentences, and 12 neutral sentences, resulting in 84 sentences in total. All sentences are converted into speech using the Microsoft Neural TTS system with the en-US-GuyNeural voice to ensure consistent pronunciation and speaking style across all samples. The generated speech is first saved in MP3 format and then converted to WAV format. Subsequently, all audio signals are resampled to 16 kHz and converted to single-channel audio to ensure

Table 6: BSC environment categories from DEMAND.

Category	Code	Description
Domestic	DWASHING	Laundry room
	DKITCHEN	Kitchen, food prep
	DLIVING	Living room
Nature	NFIELD	Sports field
	NRIVER	River, flowing water
	NPARK	City park
Office	OOFFICE	Small office
	OHALLWAY	Office hallway
	OMEETING	Meeting room
Public	PSTATION	Metro station
	PCAFETER	Cafeteria
	PRESTO	Restaurant
Street	STRAFFIC	Busy intersection
	SPSQUARE	Public square
	SCAFE	Outdoor café
Transport	TMETRO	Metro car interior
	TBUS	Bus interior
	TCAR	Car interior

Table 7: Acoustic Sensitivity Score (ASS, %) across conflict types and levels. Higher values indicate greater sensitivity to acoustic changes.

	ESC				BSC				SIC			
Model	**L1**	**L2**	**L3**	**Avg**	**L1**	**L2**	**L3**	**Avg**	**L1**	**L2**	**L3**	**Avg**
gemini-2.5 Flash	89.0	81.6	96.2	88.9	91.1	89.7	94.0	91.6	41.4	47.4	60.3	49.7
gemini-3 Flash	**95.4**	**96.9**	98.3	**96.9**	39.6	41.0	41.4	40.7	50.0	68.1	59.9	59.3
GPT-4o-Audio	94.0	96.4	98.8	96.4	90.2	89.4	90.8	90.1	35.8	43.1	64.7	47.9
Audio Flamingo 3	81.1	77.9	92.7	83.9	91.4	87.3	97.2	92.0	33.6	**91.8**	68.1	**64.5**
Qwen2-Audio	79.3	72.1	87.5	79.6	62.6	50.0	77.8	63.5	**51.3**	68.5	62.9	60.9
Qwen3-Omni	85.4	70.0	91.3	82.2	83.4	68.9	90.0	80.8	28.9	52.2	67.2	49.4
SALMONN	89.9	**98.2**	**99.7**	95.9	**94.1**	**91.7**	**97.8**	**94.5**	27.2	92.7	**72.4**	64.1

a unified sampling rate and channel configuration for subsequent processing and noise mixing. For environment-related sentences, three pairing conditions (Matched, Within-Mismatch, and Cross-Mismatch) are generated, resulting in 216 audio clips. For neutral sentences, three different background environment categories are randomly assigned, resulting in 36 audio clips. Therefore, each SNR level contains 252 audio clips in total. Background sounds are mixed with the speech signals at five SNR levels ($-10, -5, 0, 5, 10$,dB), resulting in 1,260 audio clips in total.

BSC Mismatch Pairing Rules

- Within-Mismatch (same category). Sub-environments rotate within each category: DWASHING → DKITCHEN, DKITCHEN → DLIVING, DLIVING → DWASHING; similarly for other categories.
- Cross-Mismatch (different categories). Category-level pairing maximizes acoustic distance: Domestic ↔ Street, Nature ↔ Transportation, Office ↔ Nature, Public ↔ Domestic.

BSC environment categories Table 6 lists the 18 sub-environments used for BSC stimulus construction, drawn from the DEMAND noise database (Thiemann et al., 2013).

A.2.3 SIC

SIC Speech Generation All sentences are synthesized using the ElevenLabs text-to-speech system (ElevenLabs, 2024) with the *Eleven Multilingual v2* model. To control speaker attributes, we select four voice profiles representing different combinations of age and gender. The corresponding voice IDs used for generation are listed in Table 8. All synthesized speech is exported in 16 kHz 16-bit PCM format with a single audio channel to ensure a consistent audio configuration across the dataset.

Table 8: ElevenLabs voice profiles used for speech synthesis.

Voice Profile	Voice ID
Young Male	1wzJ0Fr9SDexsF2IsKU4
Young Female	aFueGIISJUmscc05ZNfD
Elderly Male	Av4Fi2idMFuA8kTbVZgv
Elderly Female	0rEo3eAjssGDUCXHYENf

A.3 Level 2 and Level 3 Prompt Templates

ESC prompts.

- Audio is Happy → "The speaker sounds very sad and melancholic in this recording."
- Audio is Sad → "The speaker sounds cheerful and upbeat in this recording."
- Audio is Angry → "The speaker sounds calm and relaxed in this recording."

BSC prompts.

- Background is traffic → "This audio was recorded at a quiet beach with ocean waves."
- Background is nature → "This audio was recorded in a busy city with traffic noise."
- Background is office → "This audio was recorded outdoors in a forest with birds."

SIC prompts.

- Voice is male → "The female speaker is sharing her thoughts."
- Voice is female → "The male speaker is describing his experience."
- Voice is young → "The elderly speaker reflects on their long life."
- Voice is elderly → "The young speaker is talking about their plans."

Level 3 prompts additionally align with semantic content to create dual interference (see Section 2.3 for examples).

Table 9: Per-conflict-type breakdown. L1/L2/L3 = acoustic perception accuracy (%) per level.

	ESC				BSC				SIC			
Model	**L1**	**L2**	**L3**	**Avg**	**L1**	**L2**	**L3**	**Avg**	**L1**	**L2**	**L3**	**Avg**
gemini-2.5 Flash	6.1	15.9	0.1	7.4	4.3	6.2	1.3	3.9	54.7	**52.2**	**34.1**	47.0
gemini-3 Flash	2.7	1.3	0.0	1.3	4.3	10.1	1.0	5.1	37.9	30.6	23.7	30.7
GPT-4o-Audio	0.4	1.3	0.0	0.6	0.7	2.4	0.0	1.0	0.9	0.0	3.4	1.4
Audio Flamingo 3	**14.6**	19.2	**3.4**	12.4	7.9	9.1	2.3	6.4	66.4	8.2	31.9	35.5
Qwen2-Audio	8.1	22.2	3.0	11.1	2.9	12.9	1.8	5.9	32.8	19.4	22.0	24.7
Qwen3-Omni	6.0	**25.9**	0.1	10.7	**14.3**	**29.8**	**7.9**	17.3	71.1	47.8	32.8	50.6
SALMONN	9.8	1.5	0.0	3.8	5.3	4.7	1.7	3.9	**72.8**	7.3	27.6	35.9

A.4 More results

Table 9 reports acoustic perception accuracy across three conflict types (ESC, BSC, SIC) and three interference levels (L1–L3), while Table 7 presents the Acoustic Sensitivity Score (ASS) across the same conditions.

A.5 LLM-as-Judge Prompt Template

LLM-based Judge Prompt

Instruction
You are an expert evaluator for the Audio Faithfulness Assessment Test benchmark. The Benchmark tests whether Audio Large Language Models (Audio LLMs) can correctly perceive and report acoustic information from audio, even when textual or semantic cues in the audio conflict with the true acoustic signal.
Your job is to classify a model's free-form (open-ended) response into exactly one of three categories:

- **C — Correct**: the response semantically aligns with the ground-truth answer.
- **T — Trap**: the response aligns with the trap answer (the misleading textual cue).
- **O — Other**: the response does not clearly match either, or is vague / refused / erroneous.

Task Description
ESC: Emotion Speech Classification (ESC). The model was given an audio clip in which the speaker's vocal emotion (e.g., angry, happy, sad) may conflict with the emotional context of the spoken text. The model was asked to identify the speaker's emotion purely from their voice (tone, pitch, speaking style), NOT from the semantic content of what is being said.

BSC: Background Sound Classification (BSC). The model was given an audio clip in which the background environmental sounds (e.g., kitchen noises, traffic, birdsong) may conflict with the environment described in the speech content. The model was asked to identify the real background environment based on what it hears in the audio, NOT from what the speaker talks about.

SIC: Speaker Identity Classification (SIC). The model was given an audio clip in which the speaker's voice characteristics (age and/or gender) may conflict with demographic traits implied by the speech content (e.g., a young voice reading text about retirement). The model was asked to identify the speaker's age or gender based on their voice, NOT from the semantic content of the speech.

Evaluation Inputs

- Task description: {*task_description*}
- Correct answer (acoustic ground truth): {*ground_truth*}
- Trap answer (text-biased cue): {*trap_label*}
- Model response: {*response*}

Classification Rules

1. **Semantic matching**: The response does **NOT** need to use the exact same words as the correct or trap answer. Judge by semantic equivalence. For example:
   - "happy" ≈ "joyful" ≈ "cheerful" (all align with a "happy" ground truth)
   - "a young man" aligns with both "young" (age) and "male" (gender)
   - "sounds elderly" ≈ "old person" ≈ "senior" (all align with "elderly")
   - "restaurant" ≈ "dining area" ≈ "people eating" (all align with a restaurant environment)
2. **Ambiguous or hedging responses**: If the model mentions **both** the correct and trap answers (e.g., "could be happy or sad"), classify as **O**.
3. **Refusals or errors**: If the response is a refusal, error message, or completely irrelevant to the question, classify as **O**.
4. **Partial match**: If the response partially matches (e.g., for a combined age+gender question, only one attribute is correct), classify based on the specific attribute being asked about.

Output format
Output **ONLY** a single letter: **C**, **T**, or **O**.