MAKING AUDIO MATTER: MMALR FOR BENCHMARKING AUDIO LISTENING, UNDERSTANDING, AND REASONING WITH COUNTER-PRIOR CONSTRUCTION

Anonymous authors
Paper under double-blind review

ABSTRACT

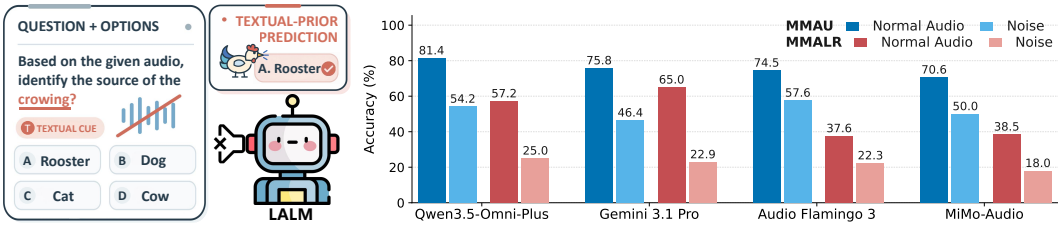
Large audio-language models (LALMs) have made substantial progress across a wide range of audio tasks. To evaluate them consistently, Audio Question Answering (AQA) has become a mainstream paradigm with a unified cross-task interface. However, existing evaluations usually rest on an implicit assumption: a model's response is taken as its analysis of the audio. But in practice, the textual content alone may induce a preference for a particular answer, which arises from the model's textual prior. The effect of the prior becomes evident when audio is replaced with random noise: a considerable portion of questions can still be correctly answered, suggesting that high accuracy on existing benchmarks may not stem from genuine audio processing capabilities. Meanwhile, whether textual priors still affect model predictions with normal audio remains unexamined. To fill this gap, we introduce MMALR, a multi-domain, multi-task benchmark for audio listening, understanding, and reasoning that uses counter-prior construction to make audio evidence play a greater role in AQA. Specifically, MMALR anchors each correct answer in audio evidence and turns plausible textual predictions into distractors, making audio indispensable for reliable answering. We systematically evaluate open- and closed-source LALMs as well as human participants. Results show that under uninformative-audio conditions, MMALR yields the lowest model accuracy among all compared benchmarks. Furthermore, paired predictions under uninformative- and normal-audio conditions on MMALR show that models differ in their textual tendencies, while most models still exhibit consistent incorrect tendencies under the normal-audio condition. MMALR thus provides a more reliable benchmark for evaluating LALMs' audio listening, understanding, and reasoning, and a basis for analyzing their text-audio modality dependency.

1 INTRODUCTION

In recent years, large audio-language models (LALMs) have made substantial progress on a wide range of audio tasks Su et al. (2026); Team (2026); Yang et al. (2026). To systematically evaluate these models, Audio Question Answering (AQA) provides a unified interface: the model receives an audio clip, a textual question, and several candidate options, and then selects the correct answer Gong et al. (2024; 2023). This format enables consistent comparison across tasks and has become a standard tool for evaluating LALMs. Recent benchmarks have further expanded AQA along audio domains, task types, and reasoning complexity Sakshi et al. (2025); Ma et al. (2025a).

However, these evaluations implicitly rest on a premise that has not been sufficiently examined: a model's response is taken as its analysis of the audio. Figure 1a gives an example from MMAU (test-mini split) Sakshi et al. (2025): some questions can be answered correctly even without audio, because the question text directly provides lexical cues related to the correct answer Glazer et al. (2026); Pang et al. (2026). In addition, probabilistic tendencies, logical associations, and other cues may allow the model to answer solely from textual information, i.e., rely on a **textual prior** Foo et al. (2026); He et al. (2026a). Figure 1b provides corresponding statistics: when the normal audio in MMAU is replaced with random noise, the evaluated LALMs still achieve 46.4%-57.6% accuracy, indicating that a considerable portion of questions can be answered without acoustic information. This exposes a limitation of existing benchmarks: such text-only solvability can inflate scores and

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) A Text-Solvable Example (b) Normal-Audio and Random-Noise Accuracy

Figure 1: **Text-Solvable Analysis on MMAU and our MMALR.** (a) An example from MMAU where the question text alone provides cues to the correct answer, allowing a model to answer correctly without listening. (b) Accuracy of LALMs with normal-audio and random-noise input. On MMAU, models can answer a considerable portion of questions from text alone, whereas on MMALR counter-prior construction leads to a marked drop in accuracy under random-noise input.

make it unclear whether a model truly relies on acoustic evidence Lee et al. (2026); Kuan & Lee (2025). The concern is particularly acute in safety-critical scenarios such as security monitoring and medical audio analysis, where a model that performs well on current benchmarks but lacks genuine listening ability may fail precisely when listening matters most.

To address this issue, we propose **MMALR**, a **Multi-domain Multi-task Audio Listening, Understanding, and Reasoning benchmark** that controls the relationship between textual cues and acoustic evidence through **counter-prior construction**. Specifically, MMALR first grounds the correct answer in identifiable, manually verified audio evidence, and then turns plausible textual predictions into distractors. This design aims to make audio truly matter in AQA benchmarks by ensuring that the correct answer cannot be reliably determined from text alone and that decisive acoustic evidence is required to resolve the ambiguity among plausible textual distractors. MMALR follows the mainstream evaluation protocol and uses accuracy under normal-audio conditions as the primary metric. By substantially reducing the cases that can be answered from text alone, counter-prior construction makes this accuracy a more reliable measure of models' overall multi-domain, multi-task AQA ability. The main contributions of this paper are as follows:

- We propose MMALR through counter-prior construction, covering speech, music, sounds, and their mixtures across four task levels with 17 fine-grained subcategories.
- We systematically evaluate mainstream LALMs and human participants under normal- and uninformative-audio conditions. Under uninformative audio, model accuracy on MMALR is the lowest among all compared benchmarks, indicating that counter-prior construction reduces the direct support available from text-only answering. Under normal audio, MMALR remains challenging for all models: the best model, Gemini 3.1 Pro, reaches 65.00%, the best open-source model, MiMo-Audio-Instruct, reaches only 38.50%, and human annotators achieve an average of 66.63%.
- We conduct paired comparisons under uninformative- and normal-audio conditions, revealing modality-dependent behaviors in LALMs: while models differ in their textual preferences, most still select the incorrect answer expected by the text when answering AQA questions, providing evidence for textual modality dependency.

2 Related Work

2.1 Large Audio Language Models

LALMs typically combine an audio encoder with an LLM: the encoder converts acoustic signals into continuous representations, which are projected into a pretrained LLM together with textual instructions, and the LLM generates the response. Representative general-purpose audio understanding models include the Qwen series Chu et al. (2024), the Audio Flamingo series Ghosh et al. (2025a; 2026), MiMo-Audio Zhang et al. (2025), and Kimi-Audio Ding et al. (2025). Some works further enhance reasoning via post-training, such as Audio-Reasoner Xie et al. (2025) and R1-AQA

Li et al. (2025). Another trend is omni-modal models that jointly handle multiple modalities, such as Qwen2.5-Omni Xu et al. (2025), and the closed-source Qwen3.5-Omni Team (2026) and Gemini Google (2026). These models achieve substantial gains on core audio tasks such as automatic speech recognition (ASR) Ma et al. (2025b), audio captioning (AC) Chen et al. (2025), and acoustic scene classification Ghosh et al. (2025b), and their strong instruction-following capabilities Xie et al. (2025) enable complex reasoning tasks to be uniformly implemented and evaluated via AQA.

Table 1: Comparison of LALMs AQA benchmarks: (i) **Domain**: sound, music, speech, or their mixtures; ✓ indicates that the mixed domain uses artificially synthesized samples. (ii) **Scope**: listening, understanding, and reasoning; ✓ in the Listening column indicates that although samples are designed to probe acoustic perception, textual information cues are not controlled, and ✓ in the Reasoning column indicates a lack of deep reasoning samples. (iii) **Counter-Prior**: counter-textual-prior design. (iv) **New Data**: newly collected data or data sourced from existing datasets; ✓ indicates partial reuse. (v) **Sample-As-A-Task**: whether each sample is evaluated as an independent task.

Benchmark	Domain				Scope			Counter-Prior	New data	Sample-As-A-Task
	Sound	Music	Speech	Mix	Listen	Understand	Reason			
AudioBench	✓	✓	×	×	✓	✓	×	×	×	×
AIR-Bench	✓	✓	✓	✓	✓	✓	✓	×	×	×
MMAU	✓	✓	✓	×	✓	✓	✓	×	✓	✓
MMAR	✓	✓	✓	✓	✓	✓	✓	×	✓	✓
MMSU	×	×	✓	×	✓	✓	✓	×	✓	✓
ADQA-Bench	✓	✓	✓	×	✓	✓	×	×	✓	✓
MMALR	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

2.2 AQA Benchmarks for Large Audio Language Models

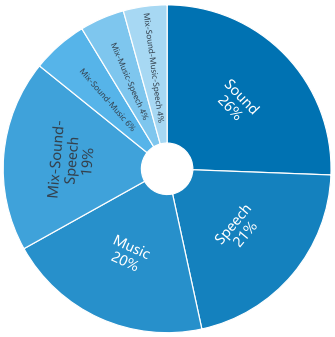
AQA has become a mainstream and general framework for evaluating LALMs. Under the multiple-choice question (MCQ) format, a model receives an audio clip, a textual question, and a set of candidate answers, and is evaluated by selecting the correct option. Compared with task-specific evaluation, AQA provides a unified interface for assessing models across different audio domains and task types. Table 1 summarizes these benchmarks along five dimensions. Here, audio listening refers to extracting question-relevant acoustic evidence; audio understanding refers to identifying the sound source, event, or object from that evidence and further interpreting its semantics; and audio reasoning refers to integrating multiple acoustic events with temporal, spatial, contextual, cultural, or professional knowledge while requiring multi-step inference. AudioBench Wang et al. (2025), AIR-Bench Yang et al. (2024), and MMAU Sakshi et al. (2025) cover multiple audio understanding tasks involving sound, music, and speech; MMAR Ma et al. (2025a) further organizes a broader set of audio reasoning tasks and extends evaluation to mixed audio scenarios; and MMSU Wang et al. (2026) focuses on speech-centered understanding and reasoning. Recently, ADQA-Bench He et al. (2026b) enhances the audio dependence of questions by filtering out those that can be answered from text alone. The comparison shows that these benchmarks provide complementary capability coverage, yet most of them do not treat the influence of textual priors as an explicit and controllable evaluation factor. In contrast, MMALR is designed for multi-domain, multi-dimensional evaluation of LALMs. Beyond covering sound, music, speech, and mixed domains, it introduces textual-prior distractors through counter-prior construction and provides fine-grained question categories, captions, and Chain-of-Thought (CoT) annotations, thereby supporting in-depth analysis of LALMs' listening, understanding, and reasoning abilities as well as their modality-dependent behavior.

3 The MMALR Benchmark

3.1 Overview

Domain Coverage. As shown in Figure 2a, MMALR spans seven audio domains: sound, music, speech, and mixed forms of sound-music, sound-speech, music-speech, and sound-music-speech.

Task Coverage. As illustrated in Figure 2b, tasks are organized into signal, perception, semantic, and cultural layers. Each layer is further subdivided into multiple subcategories, with detailed definitions and examples provided in Appendix A.



(a) Audio Domain



(b) Task Taxonomy

Statistic	Num.
Total samples	1,000
Audio Domains	7
Task Categories	4
Task Sub-Categories	17
Average (sec/words)	**Value**
Aud. Duration	19.03
Ques. Length	11.01
Ans. Length	8.11
Cap. Length	31.54
CoT Length	65.38

(c) Benchmark statistics

Figure 2: Overview of MMALR data distribution. (a) Distribution across seven audio domains: sound, speech, music, and their mixtures. (b) Task taxonomy across four layers: Signal, Perception, Semantic, and Cultural with detailed sub-categories. (c) Key statistics of MMALR.

Data Statistics. Key statistics are presented in Figure 2c. Each sample is constructed with counter-prior construction and annotated with an audio caption and a CoT. The average CoT length is approximately 65.38 words, substantially exceeding the 32.28 words of prior benchmarks Ma et al. (2025a), indicating richer and more elaborate reasoning steps and higher annotation density. All audio clips range from 5 to 30 seconds, with an average duration of approximately 19.03 seconds.

3.2 Data Curation

As shown in Figure 3, we constructed the MMALR benchmark following a five-stage framework:

Overall Design. To ensure broad task coverage and encourage genuine use of acoustic evidence in AQA evaluation, we define two complementary goals through multiple rounds of expert discussion: task coverage and audio grounding. First, we build on the four-level task taxonomy of signal, perception, semantic, and cultural Ma et al. (2025a), and adjust its finer-grained divisions into 17 subcategories, where each sample is determined by the primary type of audio evidence it requires. Second, we introduce the principle of **counter-prior construction**. Specifically, we ground every correct answer in clear and identifiable acoustic evidence, and then construct distractors from textual preferences that differ from the correct answer, together with acoustically confusable alternatives. This principle makes audio matter in AQA evaluation, while text alone is insufficient for correct prediction. This stage determines the subsequent annotation pipeline and produces the annotation guidelines. Annotator qualifications and guidelines are provided in Appendix B and C, respectively.
Audio-Grounded Annotation. Following the guidelines, human annotators search for publicly available videos and select clips with clear acoustic content. For each candidate instance, annotators construct an audio-grounded question and answer, and are required to ensure that the answer cannot be directly inferred from the question text alone and must be supported by the audio. They also provide a caption describing the audio, and a CoT explaining how the answer can be derived from the audio. To increase task diversity and difficulty, annotators may also request appropriate acoustic transformations, including cutting, concatenation, and order swapping.

Audio Preparation. Based on the annotated metadata, we crawl the corresponding audio clips from the source videos. For each candidate instance, annotators cut the audio according to the annotated timestamps and ensure that the clip contains sufficient contextual audio to support the corresponding question, after which the requested acoustic transformations are applied.

Filter and Distractor Construction. Following the audio-grounded annotation, textual preferences are elicited from both human annotators and an LLM (DeepSeek-V4-Flash Xu et al. (2026)), which receive only the question text and rely on linguistic cues and commonsense knowledge. At this stage, we discard a sample if all models and human annotators infer the correct answer from the question text alone, indicating that it is readily recoverable without audio. We then use the remaining predictions, which differ from the correct answer, to guide the manual construction of plausible textual-prior distractors. In addition, we add acoustically confusable options that reflect plausible

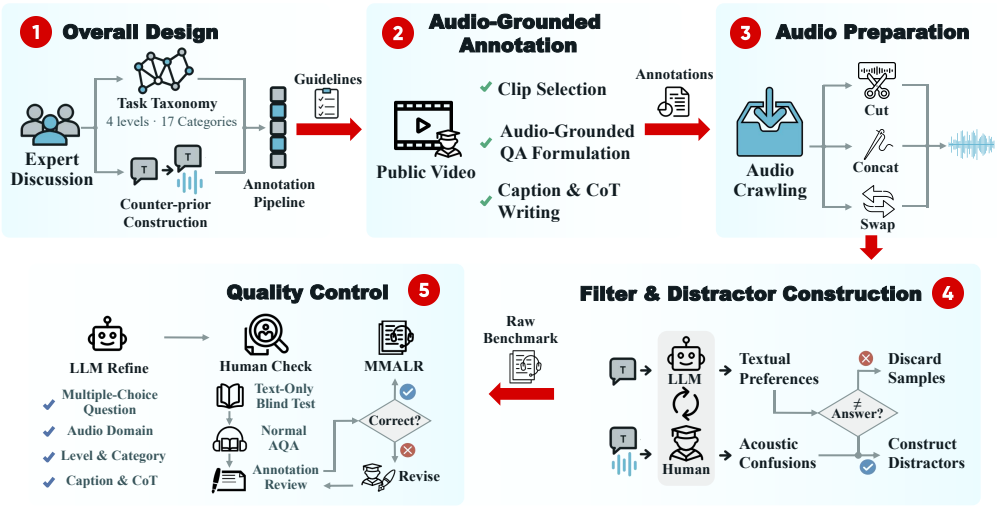


Figure 3: **A comprehensive framework for constructing the MMALR benchmark.** The framework has five stages: (1) **Overall Design** of the task taxonomy and counter-prior principle, yielding the annotation pipeline and guidelines; (2) **Audio-Grounded Annotation** from public videos, producing audio-grounded QA, captions, and CoT; (3) **Audio Preparation**, including audio crawling and optional transformations such as cutting, concatenation, and swapping; (4) **Filter and Distractor Construction**, filtering samples by human and LLM textual preferences and combining textual preferences different from the answer with acoustic confusions to design distractors; and (5) **Quality Control** via LLM-assisted refinement and three-stage human inspection.

confusions in the recording, simulating errors that models are prone to after listening to the audio. This procedure requires models to avoid relying on textual priors while also possessing sufficient audio listening, understanding, and reasoning ability to answer correctly in AQA evaluation. LLM prompts are detailed in Appendix D.

Quality Control. Finally, we conduct the following quality control steps. First, we use LLMs (DeepSeek-V4-Flash Xu et al. (2026) and GPT-5.6-Sol OpenAI (2026)) for language refinement. The refinement model receives the original question, choices, answer, audio caption, and CoT, and improves grammaticality, clarity, and consistency while preserving all semantics. Second, we develop a three-stage platform for quality inspection: a text-only blind test, a normal AQA stage, and an annotation review stage. In the first two stages, annotators answer each sample from the question and options alone, and then again after listening to the audio, assessing whether audio is necessary and whether any additional modality is required. In the third stage, annotators review the sample with the previous responses and detailed annotations, checking: (i) whether visual cues beyond the audio are required; (ii) whether the question and options are accurate and reasonable; and (iii) whether the audio caption, CoT, category, and domain are correct. Annotators can also flag other issues, such as language expression problems. Samples that fail inspection are revised and re-evaluated, and those that remain incorrect after two rounds are discarded. Only instances that pass the final validity checks are included in MMALR. LLM prompts, the annotation platform, and benchmark format examples are provided in Appendix D, E, and J, respectively.

4 Experiments and Analysis

4.1 Evaluated LALMs

We evaluate a set of recent representative LALMs, covering both open-source and proprietary closed-source models. The open-source models include: **MiMo-Audio-Instruct** Zhang et al. (2025), **Audio Flamingo 3** Ghosh et al. (2026), **Kimi-Audio-Instruct** Ding et al. (2025), **Qwen2.5-Omni-7B** and **Qwen2.5-Omni-3B** Xu et al. (2025), **R1-AQA** Li et al. (2025), **Audio-Reasoner** Xie et al. (2025), **Phi-4-Multimodal** Abouelenin et al. (2025), **Qwen2-Audio-Instruct** Chu et al.

Table 2: Accuracy (%) on MMALR under the **normal** audio condition, broken down by audio domain setting and task layer. The rightmost three columns report overall accuracy under the **normal**, **noise**, and **silent** conditions, respectively. Audio domain abbreviations: S = Sound, Sp = Speech, M = Music. Task layer abbreviations: Signal = Signal Layer, Perc. = Perception Layer, Sem. = Semantic Layer, Cult. = Cultural Layer. **Bold** denotes the best score within each group.

		Audio Domain							Task Layer				Overall		
Model	Size	Sound	Music	Speech	S-M	S-Sp	M-Sp	S-M-Sp	Signal	Perc.	Sem.	Cult.	Normal	Noise	Silent
Open-source Models															
MiMo-Audio-Instruct	7B	25.78	25.12	55.24	29.09	43.92	**63.64**	**58.14**	22.62	30.26	**51.62**	36.14	**38.50**	18.00	19.40
Audio Flamingo 3	7B	25.00	**31.53**	48.10	29.09	**44.44**	54.55	53.49	17.86	**31.84**	47.30	**39.16**	37.60	**22.30**	**23.80**
Kimi-Audio-Instruct	7B	21.48	25.12	**55.71**	10.91	31.22	47.73	44.19	22.62	20.26	46.49	36.14	32.80	16.70	17.80
Qwen2.5-Omni-7B	7B	**30.47**	28.08	50.95	27.27	35.45	43.18	41.86	**32.14**	28.68	46.49	31.93	36.10	17.20	16.60
Qwen2.5-Omni-3B	3B	25.00	28.08	44.76	23.64	32.80	59.09	39.53	29.76	26.32	41.89	31.93	33.30	16.20	16.40
R1-AQA	8.4B	29.30	22.17	39.05	25.45	31.22	40.91	32.56	22.62	26.84	34.86	34.34	30.70	18.00	16.80
Audio-Reasoner	8.4B	25.00	22.66	36.19	21.82	29.63	45.45	32.56	21.43	25.79	34.32	27.11	28.80	19.40	19.80
Phi-4-Multimodal	5.6B	19.92	22.17	43.81	23.64	26.98	43.18	41.86	19.05	20.79	37.57	33.13	28.90	18.30	19.60
Qwen2-Audio-Instruct	8.4B	25.39	22.66	34.76	21.82	24.34	34.09	32.56	19.05	24.21	29.46	32.53	27.10	16.40	16.90
GAMA	7B	26.56	19.70	18.57	**30.91**	19.05	15.91	20.93	21.43	25.26	21.62	13.25	21.60	20.10	22.20
LTU	7B	14.84	13.79	30.00	9.09	20.11	27.27	27.91	16.67	14.74	24.59	21.08	19.60	19.20	19.50
Avg.	–	24.43	23.73	41.56	22.98	30.83	43.18	38.69	22.30	25.00	37.84	30.61	30.45	18.35	18.98
Closed-source Models															
Qwen3.5-Omni-Plus	–	42.58	48.28	77.62	43.64	61.90	70.45	69.77	48.81	48.16	68.11	57.83	57.20	**25.00**	**26.50**
Qwen3.5-Omni-Flash	–	32.42	36.45	65.24	29.09	50.79	68.18	65.12	30.95	36.32	59.46	48.19	46.40	20.00	21.10
Gemini 3.1 Pro	–	**51.95**	**55.67**	**80.95**	**52.73**	**74.07**	**72.73**	**76.74**	**53.57**	**57.37**	**77.30**	**60.84**	**65.00**	22.90	22.80
Avg.	–	42.32	46.80	74.60	41.82	62.25	70.45	70.54	44.44	47.28	68.29	55.62	56.20	22.63	23.47
Human															
Annotator 1	–	**61.33**	51.72	76.67	**80.00**	76.19	**72.73**	**79.07**	**67.86**	62.37	74.32	65.06	**67.70**	–	**19.40**
Annotator 2	–	54.30	**64.04**	70.48	65.45	74.07	**72.73**	69.77	64.29	57.11	72.97	**68.67**	65.50	–	16.50
Annotator 3	–	53.52	57.64	**77.14**	74.55	**78.84**	70.45	69.77	51.19	**62.63**	**76.22**	62.65	66.70	–	14.30
Avg.	–	56.38	57.80	74.76	73.33	76.37	71.97	72.87	61.11	60.70	74.50	65.46	66.63	–	16.73

(2024) **GAMA** Ghosh et al. (2024) and **LTU** Gong et al. (2024). For closed-source models, we include **Gemini 3.1 Pro** Google (2026), **Qwen3.5-Omni-Flash** and **Qwen3.5-Omni-Plus** Team (2026).

4.2 Evaluation Protocol

For AQA evaluation, we use a unified multiple-choice prompt that instructs the model to select the correct option, structured as: "{question}\nPlease choose the answer from the following options:\n{choices_block}", where options are automatically labeled with letters. R1-AQA is evaluated with an additional instruction: "Output the final answer in `<answer> </answer>`.", as its post-training is specifically optimized for this format. We evaluate whether the model output matches the correct option. Additional implementation details are provided in Appendix H.

4.3 Overall Performance on MMALR

Table 2 reports accuracy under the normal-audio condition. MMALR remains challenging for all evaluated LALMs. Gemini 3.1 Pro achieves the strongest overall performance at 65.00%, followed by Qwen3.5-Omni-Plus at 57.20%. Among open-source models, MiMo-Audio-Instruct performs best with 38.50%, followed by Audio Flamingo 3 at 37.60% and Qwen2.5-Omni-7B at 36.10%. The 26.50% gap between Gemini 3.1 Pro and the strongest open-source model highlights substantial room for improvement in open audio models. Overall, MMALR remains challenging for all evaluated LALMs, including the strongest proprietary model.

Across the three single-domain settings, most evaluated models perform better on speech than on sound or music domain. This advantage may reflect more mature speech representations and the larger amount of speech–text supervision available during LALM training, and may also be related to differences in the required reasoning depth across samples from different audio domains. Mixed-domain questions, however, do not exhibit consistently greater difficulty than single-domain ones, indicating that model performance does not decrease monotonically with the number of audio

domains involved. At the task-layer, although signal-layer questions are grounded in relatively low-level audio evidence, they yield the lowest scores for 8 of the 14 models, whereas the semantic layer achieves the highest accuracy for 12 models. The results further suggest that although current research focuses mainly on improving the high-level and multi-step reasoning capabilities of LALMs, the low-level signal analysis stage may still be the key bottleneck constraining overall performance.

4.4 Analysis under Uninformative Audio

4.4.1 Cross-Benchmark Comparison

We evaluate two uninformative-audio settings that completely remove question-relevant information from the original audio, thereby allowing us to probe the textual priors of the models. The **noise** setting follows the experimental setup of current mainstream benchmarks and replaces each original recording with Gaussian-noise input. The **silent** setting instead replaces each recording with an all-zero waveform (see Appendix H for details).

Figure 4 compares noise-condition accuracy on MMAU, MMAR, and MMALR. For every evaluated model, MMALR yields the lowest accuracy among the three benchmarks, indicating that its correct answers rely less on textual cues when informative audio is unavailable.

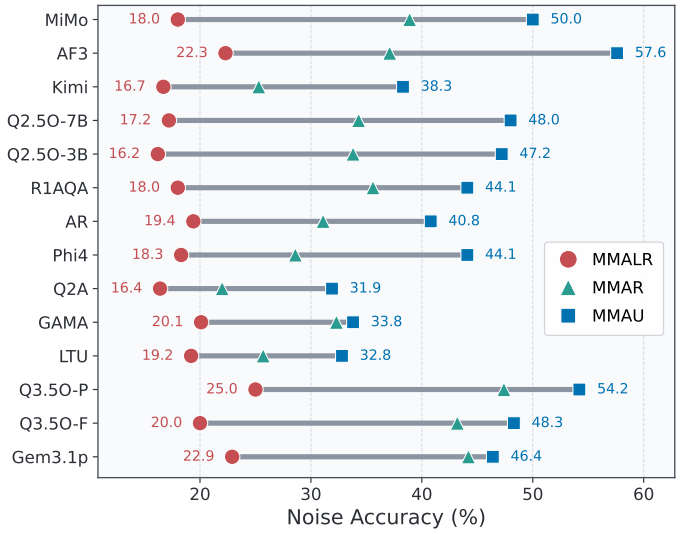


Figure 4: Noise-condition accuracy (%) on MMAU, MMAR, and MMALR. Each row is one model, and MMALR (leftmost marker) yields the lowest noise accuracy. Model abbreviations here correspond one-to-one to the full model names in Table 2; the same abbreviations apply below.

However, absolute accuracy alone does not indicate whether a model performs above chance. We therefore test whether each model's number of correct answers under noise significantly exceeds chance, using a one-sided binomial test against each benchmark's random-guessing level (25.40% for MMAU, 29.32% for MMAR, and 25.00% for MMALR), with a Bonferroni-corrected threshold of $\alpha = 0.001$. As shown in Figure 5, although MMAU and MMAR were originally reported to hover around chance level, recent models have substantially changed this picture: many now perform significantly above chance on MMAU and MMAR. In contrast, MMALR largely alleviates this issue. This supports the effectiveness of counter-prior construction in suppressing non-audio information, and suggests that different models may hold substantially different textual priors, which we examine next.

4.4.2 Model-Dependent Textual Priors

Figure 6 shows the overlap among the sets of samples answered correctly by different models under noise. The results indicate that the correct sets of different models overlap only partially; among the selected models, the intersection of samples answered correctly by all models under noise contains only 6 samples. As more models are included, this intersection is likely to shrink further. This suggests that although most samples in MMALR cannot be reliably answered from textual cues alone, different models still answer different subsets correctly, with a similar pattern on MMAU and MMAR. Therefore, textual priors are model-specific: a benchmark that appears clean for one generation of models may be re-exploited by the next through new textual priors or reasoning abilities.

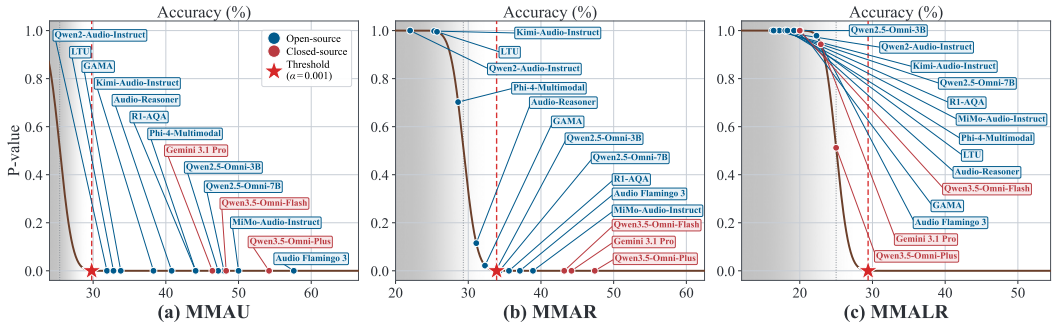


Figure 5: Right-tail p-values of noise-condition accuracy against each benchmark's chance null. Panels (a)–(c) show MMAU, MMAR, and MMALR; blue points denote open-source models and red points closed-source models. The red star marks the Bonferroni-corrected threshold $\alpha = 0.001$, with models to its right significantly above chance. Many recent LALMs exceed chance on both MMAU and MMAR, whereas MMALR largely alleviates this issue.

This also implies that multimodal evaluation should not be treated as a one-off, fixed resource, but must be continuously designed, re-examined, and updated as model capabilities evolve.

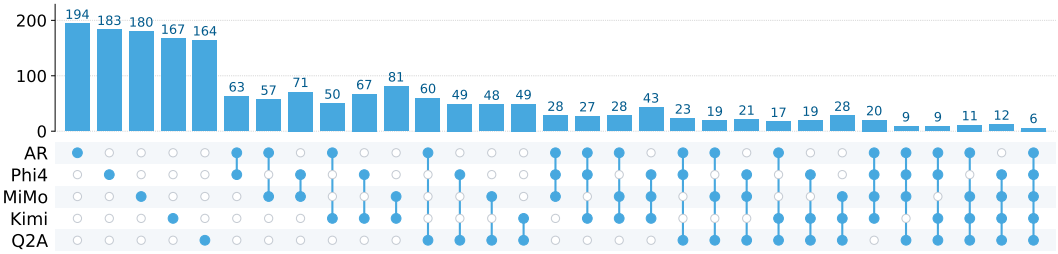


Figure 6: Overlap of correctly answered samples across models under the noise condition on MMALR. Each bar represents the number of samples correctly answered by the model subset marked by the filled dots below. Different models' correct sets overlap only partially, with few samples answered correctly by all models.

4.5 Textual Prior Errors Analysis

We next conduct paired comparisons of model predictions under uninformative- and normal-audio conditions to analyze how textual priors affect AQA behavior. For model m, we define its Textual-Prior Error (TPE) set as the samples on which the predictions under both noise and silence are identical and incorrect:

$$E_m^{\mathrm{TPE}} = \left\{ i \mid \hat{y}_{m,i}^{\mathrm{noise}} = \hat{y}_{m,i}^{\mathrm{silence}} \neq y_i \right\}, \tag{1}$$

where y_i denotes the answer. Selecting the same incorrect prediction under both uninformative conditions reflects a stable textual preference when question-relevant audio is unavailable. We define the proportion of such samples as the Prior Coverage, $\mathrm{Coverage}_m = |E_m^{\mathrm{TPE}}|/N$. On MMALR, the mean coverage across the 14 models is 64.0%. This coverage provides a sufficiently large set of samples for examining how such textual preferences affect model behavior under the normal AQA setting, where question-relevant audio is available.

For each model-sample pair in E_m^{TPE}, we compare the corresponding normal-audio prediction with the model's textual-prior error and classify it into one of three mutually exclusive outcomes:

$$\mathrm{ER}_m = \frac{\sum_{i \in E_m^{\mathrm{TPE}}} \mathbb{I}\left[\hat{y}_{m,i}^{\mathrm{normal}} = \hat{y}_{m,i}^{\mathrm{noise}}\right]}{|E_m^{\mathrm{TPE}}|}, \mathrm{AR}_m = \frac{\sum_{i \in E_m^{\mathrm{TPE}}} \mathbb{I}\left[\hat{y}_{m,i}^{\mathrm{normal}} = y_i\right]}{|E_m^{\mathrm{TPE}}|}. \tag{2}$$

Here, Error Retention (ER) refers to cases where the model's prediction under the normal AQA setting remains the same incorrect prediction as in the TPE set, and Audio Revision (AR) denotes revision to the audio-supported correct answer. The remaining cases are denoted as Other Wrong Transition (OWT), i.e., a change to another incorrect option, so that $\mathrm{OWT}_m = 1 - \mathrm{ER}_m - \mathrm{AR}_m$.

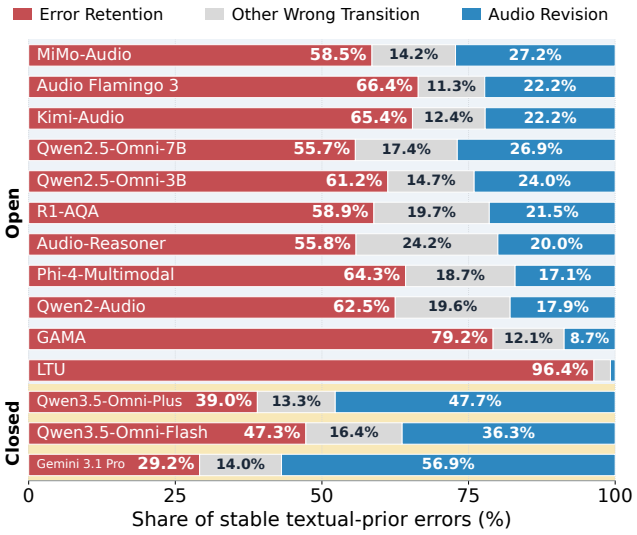


Figure 7: Normal-audio predictions on each model's TPE set. ER substantially exceeds AR for most models, indicating that models tend to retain textual-prior errors even in normal AQA. We refer to this phenomenon as Textual Prior Dominance (TPD).

Based on these definitions, we further analyze each model's behavior on its TPE set. As shown in Figure 7, ER substantially exceeds AR for most models. For open-source models, ER accounts for 55.7%-96.4% within E_m^{TPE}, whereas AR ranges only from 0.8% (LTU) to 27.2%. Closed-source models exhibit stronger but still incomplete revision behavior. Gemini 3.1 Pro and Qwen3.5-Omni-Plus are the only two evaluated models for which AR exceeds ER, yet approximately 29% and 39% of their errors, respectively, remain consistent with their textual-prior errors. These results show that, in AQA, a model's incorrect prediction under normal audio is often the same as its textual-prior error. We refer to this phenomenon as **Textual Prior Dominance (TPD)**. We suggest that TPD may arise from two cases: (i) the model's audio analysis ability is insufficient, so it fails to extract decisive acoustic evidence and falls back to its TPE; and (ii) even when the audio evidence conflicts with the textual prior, the model is still influenced by the TPE and does not revise to the audio-grounded answer. These two cases are not mutually exclusive and may co-exist to different degrees. The remaining cases, OWT, account for 2.8%-24.2% across all models, suggesting that on these samples, model predictions may be constrained by audio understanding and reasoning ability.

4.6 Comparison with Human Performance

Table 2 also reports human annotators' accuracy on MMALR under the normal-audio and silent conditions. Under normal audio, the three annotators who did not participate in data construction achieve an average overall accuracy of 66.63%, significantly outperforming most evaluated LALMs, with only a small margin over Gemini 3.1 Pro. MMALR therefore remains challenging for current models even compared with human listeners.

Under the silent condition, the annotators achieve average accuracy of only 16.73%, comparable to most models under uninformative-audio conditions, while their normal-audio accuracy is substantially higher. This contrast suggests that humans can effectively integrate textual and acoustic evidence, whereas the TPD exhibited by LALMs remains a key bottleneck.

5 Conclusion

In this paper, we introduce MMALR, a multi-domain, multi-task benchmark for audio listening, understanding, and reasoning built on counter-prior construction. By grounding each correct answer in verified audio evidence and turning plausible textual preferences into distractors, MMALR makes audio indispensable in LALM evaluation. Systematic evaluations of open- and closed-source LALMs, together with human participants, show that MMALR yields the lowest noise-condition accuracy among compared benchmarks, remains challenging for all evaluated models even relative to human listeners, and reveals a consistent phenomenon we term Textual Prior Dominance, in which models tend to select the option expected by their textual prior while neglecting relevant audio evidence. These results highlight that reliable audio listening, understanding, and reasoning is still far from solved, and that multimodal evaluation should be continuously re-examined as model capabilities and textual priors evolve.

AI USE STATEMENT

We used generative AI tools to generate part of the icons in the figures, for language polishing of the paper, and for parts of the data construction pipeline. The latter processes have been described in detail in the main text or the appendix. We did not use generative AI tools for the other steps that we have explicitly performed as manual annotation. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Data privacy and consent: All audio clips are extracted from publicly accessible, user-uploaded videos on platforms such as YouTube `https://www.youtube.com/` and Bilibili `https://www.bilibili.com/`. Each clip is shorter than 30 seconds, which is even shorter than the preview segments offered by commercial platforms such as Spotify, thereby minimizing copyright and privacy risks. No personally identifiable or sensitive user information is included.

Licensing and responsible use: The dataset is released under a CC-BY-NC license, which explicitly restricts its use to non-commercial academic research.

REPRODUCIBILITY STATEMENT

We submit the inference and evaluation code as supplementary material and will fully open-source all audio files and metadata upon completion of the review process. The main text and appendix describe the data construction pipeline, counter-prior construction, and evaluation protocol in detail. We hope these resources will support the reproduction of our results and facilitate future research on LALMs.

REFERENCES

Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, et al. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*, 2025.

Wenxi Chen, Ziyang Ma, Xiquan Li, Xuenan Xu, Yuzhe Liang, Zhisheng Zheng, Kai Yu, and Xie Chen. Slam-aac: Enhancing audio captioning with paraphrasing augmentation and clap-refine through llms. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

Xize Cheng, Dongjie Fu, Chenyuhao Wen, Shannon Yu, Zehan Wang, Shengpeng Ji, Siddhant Arora, Tao Jin, Shinji Watanabe, and Zhou Zhao. Aha-bench: Benchmarking audio hallucinations in large audio-language models. *Advances in Neural Information Processing Systems*, 38, 2026.

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

Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, et al. Kimi-audio technical report. *arXiv preprint arXiv:2504.18425*, 2025.

Leonardo Haw-Yang Foo, Chih-Kai Yang, Chen-An Li, Ke-Han Lu, and Hung-yi Lee. All that glitters is not audio: Rethinking text priors and audio reliance in audio-language evaluation. *arXiv preprint arXiv:2604.24401*, 2026.

Sreyan Ghosh, Sonal Kumar, Ashish Seth, Chandra Kiran Reddy Evuru, Utkarsh Tyagi, S Sakshi, Oriol Nieto, Ramani Duraiswami, and Dinesh Manocha. Gama: A large audio-language model with advanced audio understanding and complex reasoning abilities. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 6288–6313, 2024.

Sreyan Ghosh, Zhifeng Kong, Sonal Kumar, S Sakshi, Jaehyeon Kim, Wei Ping, Rafael Valle, Dinesh Manocha, and Bryan Catanzaro. Audio flamingo 2: an audio-language model with long-audio understanding and expert reasoning abilities. In *Proceedings of the 42nd International Conference on Machine Learning*, ICML'25. JMLR.org, 2025a.

Sreyan Ghosh, Sonal Kumar, Chandra Kiran Reddy Evuru, Oriol Nieto, Ramani Duraiswami, and Dinesh Manocha. Reclap: Improving zero shot audio classification by describing sounds. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025b.

Sreyan Ghosh, Arushi Goel, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang-gil Lee, Chao-Han Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, et al. Audio flamingo 3: Advancing audio intelligence with fully open large audio language models. *Advances in Neural Information Processing Systems*, 38:41819–41886, 2026.

Neta Glazer, Lenny Aharon, and Ethan Fetaya. Are audio-language models listening? audio-specialist heads for adaptive audio steering. *arXiv preprint arXiv:2603.06854*, 2026.

Yuan Gong, Alexander H Liu, Hongyin Luo, Leonid Karlinsky, and James Glass. Joint audio and speech understanding. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 1–8. IEEE, 2023.

Yuan Gong, Hongyin Luo, Alexander Liu, Leonid Karlinsky, and James R Glass. Listen, think, and understand. In *International Conference on Learning Representations*, volume 2024, pp. 18516–18545, 2024.

Google. Gemini 3.1 pro, 2026. URL https://docs.cloud.google.com/gemini-enterprise-agent-platform/models/gemini/3-1-pro.

Haolin He, Xingjian Du, Renhe Sun, Zheqi Dai, Yujia Xiao, Mingru Yang, Jiayi Zhou, Xiquan Li, Zhengxi Liu, Zining Liang, et al. Measuring audio's impact on correctness: Audio-contribution-aware post-training of large audio language models. In *International Conference on Learning Representations*, volume 2026, pp. 110411–110432, 2026a.

Haolin He, Renhe Sun, Zheqi Dai, Xingjian Du, Chunyat Wu, Zining Liang, Zhengxi Liu, Jiahe Lei, Runbang Wang, Jiayi Zhou, et al. Summary of dcase 2026 task 5: Audio-dependent question answering. *arXiv preprint arXiv:2607.18718*, 2026b.

Chun-Yi Kuan and Hung-yi Lee. Teaching Audio-Aware Large Language Models What Does Not Hear: Mitigating Hallucinations through Synthesized Negative Samples . In *Interspeech 2025*, pp. 2073–2077, 2025. doi: 10.21437/Interspeech.2025-324.

Sonal Kumar, Šimon Sedláček, Vaibhavi Lokegaonkar, Fernando López, Wenyi Yu, Nishit Anand, Hyeonggon Ryu, Lichang Chen, Maxim Plička, Miroslav Hlaváček, et al. Mmau-pro: A challenging and comprehensive benchmark for holistic evaluation of audio general intelligence. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 22688–22697, 2026.

Dongwook Lee, Sangkwon Park, Eunwoo Song, Che Hyun Lee, Youngho Cho, Junho Kim, June Young Yi, Heeseung Kim, and Sungroh Yoon. Hear who said what: Unlocking speaker-attributed reasoning via counterfactual voice grounding. *arXiv preprint arXiv:2608.29120*, 2026.

Gang Li, Jizhong Liu, Heinrich Dinkel, Yadong Niu, Junbo Zhang, and Jian Luan. Reinforcement learning outperforms supervised fine-tuning: A case study on audio question answering, 2025. URL https://arxiv.org/abs/2503.11197.

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, Tianrui Wang, Wang Yuping, Yuxuan Wang, Yihao Wu, Guanrou Yang, Jianwei Yu, Ruibin Yuan, Zhisheng Zheng, Ziya Zhou, Haina Zhu, Wei Xue, Emmanouil Benetos, Kai Yu, Eng-Siong Chng, and Xie Chen. Mmar: A challenging benchmark for deep reasoning in speech, audio, music, and their mix. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*,

volume 38, Main Conference. Curran Associates, Inc., 2025a. doi: 10.52202/085713-2252. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/610a7d6507d55be70c6d13d0b663227d-Paper-Datasets_and_Benchmarks_Track.pdf.

Ziyang Ma, Guanrou Yang, Yifan Yang, Zhifu Gao, Jiaming Wang, Zhihao Du, Fan Yu, Qian Chen, Siqi Zheng, Shiliang Zhang, et al. Speech recognition meets large language model: Benchmarking, models, and exploration. In *Proceedings of the AAAI conference on artificial intelligence*, pp. 24840–24848, 2025b.

OpenAI. Gpt 5.6 sol, 2026. URL https://openai.com/index/gpt-5-6/.

Jiacheng Pang, Ashutosh Chaubey, and Mohammad Soleymani. Do audio llms listen or read? analyzing and mitigating paralinguistic failures with voxparadox. *arXiv preprint arXiv:2605.27772*, 2026.

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

Yi Su, Jisheng Bai, Qisheng Xu, Kele Xu, and Yong Dou. Audio-language models for audio-centric tasks: A systematic survey, 2026. URL https://arxiv.org/abs/2501.15177.

Qwen Team. Qwen3.5-omni technical report. *arXiv preprint arXiv:2604.15804*, 2026.

Bin Wang, Xunlong Zou, Geyu Lin, Shuo Sun, Zhuohan Liu, Wenyu Zhang, Zhengyuan Liu, AiTi Aw, and Nancy Chen. 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*, pp. 4297–4316, 2025.

Dingdong Wang, Junan Li, Jincenzi Wu, Dongchao Yang, Xueyuan Chen, Tianhua Zhang, and Helen Meng. Mmsu: A massive multi-task spoken language understanding and reasoning benchmark. In *International Conference on Learning Representations*, volume 2026, pp. 31374–31410, 2026.

Zhifei Xie, Mingbao Lin, Zihang Liu, Pengcheng Wu, Shuicheng Yan, and Chunyan Miao. Audio-reasoner: Improving reasoning capability in large audio language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 23840–23862, 2025.

Jiaqi Xiong, Yunjia Qi, Qi Cao, Yu Zheng, Yutong Zhang, Ziteng Wang, Ruofan Liao, Weisheng Xu, and Sichen Liu. Deaf: A benchmark for diagnostic evaluation of acoustic faithfulness in audio language models. *arXiv preprint arXiv:2603.18048*, 2026.

Anyi Xu, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chenchen Ling, et al. Deepseek-v4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*, 2026.

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. Qwen2.5-omni technical report, 2025. URL https://arxiv.org/abs/2503.20215.

Chen Yang, Chufan Yu, Hanfu Chen, Jie Zhu, Jingqi Chen, Ke Chen, Wenxuan Wang, Yang Wang, Yaozhou Jiang, Yi Jiang, et al. Moss-audio technical report. *arXiv preprint arXiv:2606.01802*, 2026.

Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, Chang Zhou, et al. Air-bench: Benchmarking large audio-language models via generative comprehension. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 1979–1998, 2024.

Dong Zhang, Gang Wang, Jinlong Xue, Kai Fang, Liang Zhao, Rui Ma, Shuhuai Ren, Shuo Liu, Tao Guo, Weiji Zhuang, et al. Mimo-audio: Audio language models are few-shot learners. *arXiv preprint arXiv:2512.23808*, 2025.

Feiyu Zhao, Yiming Chen, Wenhuan Lu, Daipeng Zhang, Xianghu Yue, and Jianguo Wei. Halluaudio: A comprehensive benchmark for hallucination detection in large audio-language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 38797–38816, 2026.

A Task Definitions

To support multi-dimensional analysis of audio capabilities, MMALR organizes its tasks into a four-layer taxonomy based on the primary abstraction level of the decisive audio evidence: the Signal Layer, the Perception Layer, the Semantic Layer, and the Cultural Layer. Each layer is further divided into sub-categories that specify the type of evidence most central to resolving the sample.

A.1 Signal Layer

The Signal Layer focuses on direct analysis of low-level acoustic properties, including acoustic quality, signal changes, temporal irregularities, and other physical characteristics of the waveform or recording. These tasks typically require precise discrimination of acoustic features. It contains three sub-categories.

Sub-Category	Metadata
Acoustic Quality Analysis: Analysis of recording or signal quality, including bitrate, distortion, clarity, noise, reverberation, clipping, or other acoustic-quality differences.	**Question:** Which encoding detail is identified for the music in this recording? **Options & Answer:** **A. It was encoded at a bitrate of 16 kbps.** B. It was encoded at a bitrate of 32 kbps. C. It was encoded at a bitrate of 64 kbps. D. It was encoded at a bitrate of 128 kbps. **CoT:** A single piece of music can be heard playing. The overall sound is muddy with severe loss of high frequencies, accompanied by noticeable compression distortion and significant loss of detail. The audio is identified as being compressed at a low bitrate of only 16 kbps. Therefore, the encoding detail is a bitrate of 16 kbps.
Anomaly Detection: Identification of an acoustically unusual event, property, or configuration that deviates from the expected signal or event pattern.	**Question:** What accounts for the distinct acoustic qualities of the dripping sound? **Options & Answer:** A. The liquid is highly viscous and flowing at a slow rate. **B. The recording has been processed at a reduced playback speed.** C. The droplets are striking a hollow and highly resonant object. D. The audio was captured using a submerged underwater microphone. **CoT:** The dripping sound is audible, but the intervals between drops are noticeably elongated. The impact reverberations are prolonged, and the pitch is lower than typical water droplets. These characteristics suggest the audio has been altered, confirming that the recording has been processed at a reduced playback speed.

Sub-Category	Metadata
Audio Difference Analysis: Comparison of two sounds, segments, or repeated events to identify a difference in acoustic properties, source characteristics, or signal behavior.	**Question:** Which identification of the steel plates matches the first and second knocking sequences? **Options & Answer:** A. The first plate is sound-deadened, and the second plate is standard. B. The first plate is standard, and the second plate is also standard. C. The first plate is sound-deadened, and the second plate is also sound-deadened. **D. The first plate is standard, and the second plate is sound-deadened.** **CoT:** Two distinct types of knocking sounds are heard. The initial impacts are bright, resonant, and sustained with a higher volume, while the subsequent impacts are dull, short, and decay rapidly with a lower volume. Comparing the resonance, sustain, and volume reveals that the first plate is standard and the second plate is sound-deadened.

A.2 Perception Layer

The Perception Layer primarily focuses on the interpretation of perceptual structures in audio, such as sound sources, spatial relations, temporal sequences, event counts, acoustic correlations, musical performance and style, and environmental scenes. These tasks require the model to organize acoustic observations into perceptual objects, events, or relations. It contains six sub-categories.

Sub-Category	Metadata
Spatial Analysis: Inference of spatial relations, movement direction, distance, or source–listener configuration from acoustic cues.	**Question:** Which description best characterizes the sound source's spatial movement over time? **Options & Answer:** A. It stays near the center while its loudness varies B. It travels smoothly around the listener in a circle C. It progresses steadily from the left side to the right **D. It jumps repeatedly among positions in several directions** **CoT:** A sharp, intermittent sound is heard, with the source position continuously switching between up, down, left, and right. Therefore, it jumps repeatedly among positions in several directions.
Temporal Analysis: Interpretation of temporal order, duration, repetition, simultaneity, or event progression in the recording.	**Question:** How much time passes between each change from one instrument to the next? **Options & Answer:** A. Approximately 4 seconds B. Approximately 1 second C. Approximately 3 seconds **D. Approximately 2 seconds** **CoT:** The audio contains four distinct instrument sounds played in sequence. The violin sound is heard from 0 to 2 seconds, the trumpet from 2 to 4 seconds, the piano from 4 to 6 seconds, and the guitar from 6 to 8 seconds. By calculating the time between each change of instrument, it is determined that approximately 2 seconds pass before the next one starts. Therefore, the answer is D. Approximately 2 seconds.

Sub-Category	Metadata
Correlation Analysis: Reasoning about the relationship between multiple auditory events, such as whether one sound triggers, follows, accompanies, or explains another.	**Question:** What does the man's claim about feeling the bowl's color ultimately reveal? **Options & Answer:** A. He has developed enhanced tactile senses to compensate for his condition. **B. He is deceptively pretending to be visually impaired to manipulate others.** C. He is employing a clever tactic to verify the shop owner's honesty. D. He is performing a rehearsed comedic routine for the store's patrons. **CoT:** The question asks what the man's claim about feeling the bowl's color ultimately reveals. Based on the audio, the man claims that losing one sense makes the other senses extremely sharp, using this to explain his ability to determine the bowl's color by touch. However, color is a visual property and does not have physical characteristics that can be distinguished through touch. This indicates that his claim about heightened senses is fabricated. Therefore, he is deceptively pretending to be visually impaired to manipulate others.
Counting and Statistics: Counting repeated events or estimating numerical properties from acoustic patterns, repetitions, durations, or frequencies.	**Question:** How many complete scoop-forming motions occur while the fruit is being prepared? **Options & Answer:** A. 12 motions **B. 13 motions** C. 14 motions D. 15 motions **CoT:** The audio contains a repeating sound pattern associated with a scooping action. Each complete cycle consists of a tool being inserted into fruit flesh, rotated, and the flesh separating from the rind. Counting the number of times this full cycle repeats yields exactly 13 motions. Therefore, the total number of complete scoop-forming motions is 13.
Music Theory and Aural Skills: Identification or interpretation of musical structures and theoretical properties, such as period, melody, harmony, rhythm, tonality, mode, form, texture, or voice leading, together with aural skills for recognizing them.	**Question:** Which time signature fits the excerpt's recurring beat pattern? **Options & Answer:** A. 3/4 time **B. 5/4 time** C. 4/4 time D. 6/8 time **CoT:** The question asks which time signature fits the excerpt's recurring beat pattern. The strong beats and rhythmic accents indicate a recurring pattern of five beats per measure, though the meter is not distinctly perceptible. This corresponds to a 5/4 time signature.
Environmental Perception and Reasoning: Recognition and interpretation of environmental scenes, activities, or interactions from multiple sounds and their contextual relations.	**Question:** What was the man trying to do at the glass? **Options & Answer:** A. Test whether people outside could see him through it. **B. Play tricks on people outside while staying hidden behind it.** C. Check whether he could see people outside clearly through it. D. Get the attention of people outside from behind it. **CoT:** The woman mentions that she changed the glass so it can be seen through from the inside but not from the outside. Then, a sound of sliding on glass is heard, indicating the man is doing something on the glass. The woman says she only replaced the glass upstairs. This suggests the man is trying to take advantage of the one-way glass to play tricks on people outside while believing he is hidden. Therefore, the answer is B.

A.3 SEMANTIC LAYER

The Semantic Layer is defined broadly as the interpretation of meaning conveyed by the audio. It is not restricted to linguistic or speech semantics. This layer includes the content of an utterance or event, the emotion or intention expressed by a speaker or sound, the identification or interpretation of speakers and sound sources, and the meaning of musical structures and stylistic properties. The decisive evidence may come from speech, paralinguistic properties, non-speech events, music, or their combination. It contains four sub-categories.

Sub-Category	Metadata
Content Analysis: Extraction and interpretation of the explicit content, event, or message conveyed by speech or other meaningful audio.	**Question:** What best explains the girl's shift from answering $1 to giving $20? **Options & Answer:** A. She interpreted the figures as separate prices B. She added an extra charge to the total C. She repeated a number suggested by the situation **D. She miscalculated the costs** **CoT:** The mother asks the daughter a simple addition question: 50 cents + 50 cents = ?. The daughter answers $1, though her tone suggests she is guessing. The mother then asks, "How much do you make if it costs $1 and you sell it for $1?", and the daughter excitedly guesses $20. The mother confirms that the daughter cannot perform the required subtraction, so the shift to $20 reflects a miscalculation rather than an added charge or separate price.
Emotion and Intention: Inference of affective state, communicative intention, or socially meaningful purpose from vocal, musical, or environmental cues.	**Question:** What mismatch in the reply most plausibly precedes the first girl's cooler delivery? **Options & Answer:** A. A playful-sounding remark is met with a literal confirmation **B. A shirt-focused compliment is met with a broad declaration of affection** C. An affectionate statement is met with praise directed at clothing D. A direct request is met with a subdued sign of agreement **CoT:** Girl A gives a specific compliment about Girl B's shirt. Girl B responds with the general expression of affection, "I love you too," instead of acknowledging the specific compliment. After hearing this reply, Girl A's tone becomes noticeably colder and lower, indicating disappointment. The disappointment likely arises because the shirt-focused compliment was met with a broad declaration of affection rather than a direct acknowledgment or reciprocal comment about the shirt. Therefore, the first girl's cooler delivery is best explained by a shirt-focused compliment being met with a broad declaration of affection.
Speaker Analysis: Analysis of speaker identity-related properties, accent, pronunciation, paralinguistic behavior, dialogue roles, or speaker interaction.	**Question:** What best explains the father's sharp inhalation as the son continues speaking? **Options & Answer:** A. The son's account makes the father feel frustrated by the dispute **B. The son's increasingly complex English is difficult for the father to follow** C. The son's account makes the father realize a painful truth D. The son's explanation gives the father unexpected news **CoT:** The father initially asks a question using poor Chinglish, and a confused sound effect indicates that the son does not understand. After the father switches to Chinese, the son immediately comprehends and starts explaining in complex, fluent English. As the son's speech becomes longer and more fluent, the father produces a sharp inhaling sound. This suggests difficulty following the increasingly complex English, so the son's English is difficult for the father to follow.

Sub-Category	Metadata
Music Structure and Style: Identification or interpretation of musical structures and stylistic properties, such as period, melody, harmony, rhythm, tonality, mode, form, texture, voice leading, instrumentation, performance techniques, dynamics, articulation, ensemble interaction, and stylistic features.	**Question:** Which compositional techniques are used in this piano excerpt to create its intense, unsettled character? **Options & Answer:** A. Unresolved dominant chords and whole-tone scales B. Sustained tone clusters and gradual dynamic swells **C. Consecutive dissonant chords and chromatic intervals** D. Continuous chromaticism with no clear tonal center **CoT:** The question asks about the compositional techniques used in this piano excerpt to create its intense, unsettled character. Listening to the audio, the chords are dissonant and appear in succession without resolving to consonant intervals. Additionally, chromatic intervals are frequently used in the melody, further intensifying the unease and morbid quality. These auditory cues align with the option describing consecutive dissonant chords and chromatic intervals. Therefore, the correct answer is: Consecutive dissonant chords and chromatic intervals.

A.4 Cultural Layer

The Cultural Layer requires the integration of audio evidence with higher-level contextual, cultural, aesthetic, imaginative, or professional knowledge. These tasks may involve interpreting socially situated behavior, evaluating artistic properties, inferring an imagined process, or applying domain-specific knowledge to an acoustic observation. It contains four sub-categories.

Sub-Category	Metadata
Cultural and Contextual Reasoning: Interpretation that requires social, cultural, situational, or pragmatic knowledge beyond the directly perceived acoustic event.	**Question:** What idea underlies the woman's response to the man's request? **Options & Answer:** A. A social convention favoring formal introductions before further contact B. A cultural preference for speaking in person instead of calling C. A practical concern about using telephone numbers across national borders **D. A historical practice of replacing personal identities with assigned numbers** **CoT:** The man is identified as German and the woman as Jewish based on the dialogue. The woman's statement about Jews using names instead of numbers references the historical practice of Nazis replacing Jewish identities with concentration camp numbers. In English, "phone number" and "number" share the same word. The woman uses this historical allusion as a metaphorical refusal to provide her phone number. This aligns with the option describing a historical practice of replacing personal identities with assigned numbers.
Imagination: Inference about an imagined, abstract, simulated, or counterfactual process represented by the audio.	**Question:** What imagery is this performance most likely intended to evoke? **Options & Answer:** A. A tranquil bamboo grove enveloped in a thick, swirling morning mist B. A quiet courtyard bathed in pale moonlight, with water gently reflecting the night sky **C. A vast desert landscape with ruins emerging from sand and wind** D. A secluded coastal cliff overlooking a calm, vast, and endless ocean **CoT:** The melody, timbre, and playing techniques in this trio are analyzed. The texture of the pipa, the piano accompaniment, and the use of slides and harmonics create a sense of openness and distance. The modal quality evokes a Central Asian character. These auditory elements collectively suggest a vast desert landscape with ancient ruins. Therefore, the correct answer is: C. A vast desert landscape with ruins emerging from sand and wind.

Sub-Category	Metadata
Aesthetic Analysis: Evaluation of artistic, expressive, stylistic, or aesthetic properties of music, speech, or other sounds.	**Question:** How are the solos of top performers? **Options & Answer:** A. Emotionally rich B. Use of unconventional techniques C. Highly impressive **D. Occasional mishaps** **CoT:** The audio features a clarinet solo over a string accompaniment. During the solo, a sudden sharp and distorted sound is heard. Based on the characteristics of the clarinet and the nature of the sound, this can be identified as a playing mishap or cracked note. Therefore, even top performers may have occasional mishaps.
Professional Knowledge and Reasoning: Application of specialized knowledge from medicine, engineering, sports, science, or other professional domains to interpret the audio.	**Question:** What might this heartbeat sound indicate? **Options & Answer:** A. Aortic valve stenosis B. The heart is pumping blood normally **C. Ventricular septal defect** D. Mitral valve prolapse **CoT:** The heartbeat sounds are audible, with a holosystolic harsh murmur present during the systolic phase. Based on medical auscultation knowledge, this is a typical heart sound manifestation of ventricular septal defect. Therefore, the heartbeat sound indicates ventricular septal defect.

B Annotator Team Qualifications

Annotation, correction, and quality assurance for MMALR are carried out by a team with backgrounds in computer science and artificial intelligence, whose research directions cover directly relevant areas such as speech processing, sound signal processing and general audio understanding, and multimodal learning. Among them, two members have extensive experience in large-scale data annotation and are familiar with annotation guideline development, quality control, and consistency checking, and are responsible for organizing and overseeing the annotation work.

For music-related tasks, two team members have more than ten years of music learning background and are able to make relatively professional judgments on pitch, rhythm, harmony, mode, form, instrumental timbre, and performance style. To further ensure the professionalism of music-related tasks, the team also invited external experts as consultants, including a professional musician and two experts with signal processing backgrounds, who participate in the review of annotation guidelines and the re-examination of difficult samples.

C Audio-Grounded Annotation Guidelines

This document provides guidelines for the initial annotation of audio-grounded AQA samples. Annotators are required to complete audio clip selection, question and answer construction, task category assignment, audio captioning, Chain-of-Thought (CoT) writing, and textual preference recording.

C.1 Source and Audio Requirements

1. Audio should preferably be sourced from real-world recordings or publicly available videos, including platforms such as YouTube, Bilibili, or other public sources. Annotators must record the source link and the start and end timestamps of the selected clip.
2. Each audio clip should be 5–30 seconds in duration and should contain sufficient context.
3. The selected clip must contain unique or decisive acoustic evidence sufficient to answer the question. The decisive evidence must be present within the audio clip itself, and must not rely on visual frames, titles, subtitles, uploader descriptions, or other external metadata.

4. Audio quality should be suitable for human listening, avoiding extremely low volume, severe clipping, excessive reverberation, severe compression distortion, or persistent strong noise.
5. Audio may be trimmed, concatenated, or temporally reordered, provided that it still contains clear and verifiable decisive acoustic evidence.

C.2 Question and Answer Requirements

1. Questions should be constructed based on the audio evidence and the intended task category, asking about properties of the recording that are observable, inferable, or require domain knowledge for judgment.
2. The answer must depend on unique or decisive auditory information in the audio and must not be reliably determinable from the question text alone. This "not determinable from text" property should arise from insufficient information, rather than from constructing a direct contradiction between the text and the audio facts.
3. Questions must not be answerable solely by relying on ASR transcriptions. If the answer can be obtained simply by reading the speech content, without requiring other acoustic cues or further reasoning, the sample should not be retained.
4. Answers should be concise and precise, and should map clearly to a single candidate answer. Avoid unnecessary proper nouns, source-specific identifiers, subtitles, or other metadata that may reveal the answer.

C.3 Audio Domain Annotation

Each sample must be annotated with exactly one primary audio domain according to its main audio content: sound, music, speech, mix-sound-music, mix-sound-speech, mix-music-speech, or mix-sound-music-speech. If the audio contains multiple salient audio modalities, the corresponding mixed domain should be preferred. If a domain appears only briefly and is irrelevant to the question, it may be ignored, and the primary domain should be annotated according to the content most relevant to the question. For samples containing speech, the language must be annotated; for samples without speech, the language field should be recorded as null. Languages should use explicit names, such as English, Chinese, Korean, or multiple languages.

C.4 Task Layer and Sub-category Assignment

Assign each sample one primary task layer and one primary sub-category according to the type of evidence that is most decisive for distinguishing the answer from other possible answers. The task layer represents the primary abstraction level of the decisive audio evidence and is not a difficulty ranking. Use the following four task layers and their corresponding sub-categories:

- **Signal Layer**: direct analysis of low-level acoustic properties, signal changes, audio quality, or anomalies. Sub-categories include Acoustic Quality Analysis, Anomaly Detection, and Audio Difference Analysis.
- **Perception Layer**: perceptual interpretation of sound sources, events, temporal or spatial relations, counts, musical structures, or environmental scenes. Sub-categories include Spatial Analysis, Temporal Analysis, Correlation Analysis, Counting and Statistics, Music Theory and Aural Skills, and Environmental Perception and Reasoning.
- **Semantic Layer**: broad interpretation of the meaning conveyed by audio, including content, emotion, intention, and speaker-related information. This layer is not limited to linguistic semantics. Sub-categories include Content Analysis, Emotion and Intention, Speaker Analysis, and Music Structure and Style.
- **Cultural Layer**: integration of audio evidence with cultural, contextual, aesthetic, imaginative, or professional knowledge. Sub-categories include Cultural and Contextual Reasoning, Imagination, Aesthetic Analysis, and Professional Knowledge and Reasoning.

If a sample fits multiple categories, the category corresponding to the evidence that is most decisive for distinguishing the answer should be selected. Alternative categories should be recorded when necessary.

C.5 Self Quality Inspection

Each candidate sample undergoes the following inspection stages.

1. **Text-only blind test.** Inspectors answer the sample using only the question and candidate options. They record whether the answer is directly recoverable and identify the most plausible textual preference.
2. **Audio question-answering test.** Inspectors listen to the recording and answer the same sample. They verify that the answer is supported by identifiable audio evidence and that the relevant evidence is present within the selected clip.

C.6 Required Annotation Fields

For each sample, annotators should provide the source identifier, source URL, and timestamp; the audio path together with any preprocessing or transformation information; the question text; the answer; the audio caption; the CoT; the task layer and sub-category; the audio domain; and the human textual preference.

D Large Language Model Implementation Details

D.1 Counter-Prior Construction

For each question, we elicit the textual preference that a model would form without access to the audio. The model receives only the question text, does not use the audio or the answer, and is instructed to give a short answer. For each question, each model family is called exactly once when the call succeeds. The system prompt used is shown in Listing 1.

Listing 1: Prompt for eliciting textual preferences.

```
system:
You are a helpful assistant for general knowledge questions.
Answer directly with a short, specific phrase (2-10 words).
Base your answer on common sense.

user:
{question}
```

For some models on some samples, the raw answer may be a refusal, e.g., claiming that audio input is required. Such raw answers are not included as references for distractor construction.

D.2 LLM-Assisted Language Refinement

The audio caption and CoT are refined by two text-only LLMs, DeepSeek-V4-Flash and GPT-5.6-Sol. We choose these two models because they do not belong to the same model families as the LALMs to be evaluated. The models receive the original audio content description, the original CoT, and the reference question-answer pair. DeepSeek-V4-Flash first returns a refined description and CoT, after which GPT-5.6-Sol makes any necessary modifications across the full content. The prompts used are shown in Listing 2 and Listing 3, respectively. The samples refined in this stage are subsequently rechecked by human annotators.

Listing 2: Prompt1 for LLM-assisted language refinement.

```
system:
You are a professional audio content description and reasoning refinement
expert. Your task is to polish the 'caption' (audio content description)
and 'thinking' (reasoning process) based on the provided audio metadata.
```

```
Core rules:
1. Distinguish the tasks: 'caption' is an objective description of audio
    events; 'thinking' is the step-by-step reasoning that answers the
    question based solely on auditory cues, the question, and the options
    -it must not start from events that were not explicitly heard.
2. All output must be in English.
3. Strictly stay within the given information; do not add speculation or
    extra details.
4. Terminology consistency: align wording with 'question' and 'answer';
    the final conclusion of 'thinking' must be the content of 'answer'.
5. Output format: return only a JSON object with the keys '
    refined_caption' and 'refined_thinking'.

user:
{question}
{options}
{answer}
{caption}
{thinking}
```

Listing 3: Prompt2 for LLM-assisted language refinement.

```
system:
You are a benchmark editor in the field of audio question answering (AQA)
    . Your task is to polish the language of multiple-choice samples so
    that they are clearer, more natural, and more consistent, while
    preserving the original meaning. Without changing the gold fact, you
    should also try to avoid the question or the options themselves
    directly revealing the answer.

Operational constraints:
1. Do not fabricate any events, objects, lyrics, or dialogue that are not
     present in the caption / thinking.
2. The rewritten correct answer must refer to the same fact as the
    original answer. Option order may be changed, but the fact
    corresponding to the correct answer must not be changed.
3. All output must be in English.
4. The number of options must be the same as in the original sample; each
     option should be given in the form A. / B. / C. / D.
5. All options should be roughly comparable in length, specificity,
    terminology density, and syntactic style.
6. The question must not reveal the answer, must not restate the caption,
     and must not include vague filler such as "according to the audio /
    listen carefully".
7. Polish the thinking as well: it must rely on the original thinking and
     be minimally modified only according to the new question / choices /
     answer.
   - Preserve the auditory evidence, reasoning steps, and conclusion
    facts already confirmed as correct in the original thinking.
   - Fix any inconsistency with the new question wording, new option
    wording, or new answer letter / full text.
   - Do not write polishing explanations into the thinking.
8. Do not output reasoning, analysis, markdown, or any extra fields.

Return only the following JSON object:
{"question": "...", "choices": ["A. ...", "B. ...", "C. ...", "D. ..."],
    "answer": "X. ...", "thinking": "..."}
where the answer must be exactly identical to the unique correct item in
    choices.

user:
{question}
{options}
{answer}
```

```
{caption}
{thinking}
```

E Data Annotation Platform

We develop an annotation platform that supports multimodal presentation, multi-stage review, and both structured and unstructured feedback, with which we complete the iterative quality inspection step. All fields of each sample, including the question, options, answer, caption, CoT, category, domain, and other metadata, are reviewed and corrected through this system; an overview of the platform interface is shown in Figure 8. In addition, we use the same platform to complete the Human Evaluation step, where annotators can answer questions without audio and then perform the normal audio-based question answering.

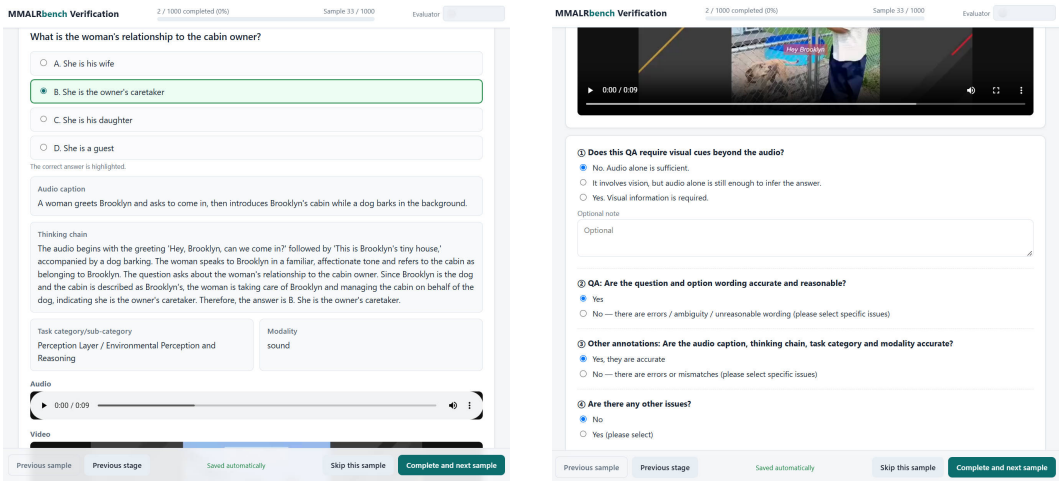


Figure 8: The annotation platform used for iterative quality inspection and human evaluation.

F Benchmarking Candidates

F.1 Open-source Models

MiMo-Audio-Instruct. Developed by Xiaomi, MiMo-Audio-Instruct is the instruction-tuned variant of MiMo-Audio, a unified audio-language model pre-trained on over 100 million hours of speech data. Open-source with weights and code.

Audio Flamingo 3. Developed by NVIDIA and UMD, Audio Flamingo 3 extends the Audio Flamingo series with the AF-Whisper unified audio encoder and a five-stage training curriculum. It supports long-audio understanding and reasoning up to 10 minutes across speech, sound, and music. Open-source with weights and code.

Kimi-Audio-Instruct. Developed by Moonshot AI, Kimi-Audio is an open-source audio foundation model for understanding, generation, and conversation. The Instruct checkpoint employs a hybrid audio input of continuous acoustic and discrete semantic tokens, with a dual-head architecture for parallel text and audio generation. Open-source with weights and code.

Qwen2.5-Omni-7B. Developed by Alibaba, Qwen2.5-Omni-7B is a dual-path "Thinker-Talker" model that supports real-time multimodal interaction over text, image, audio, and video. It handles streaming inputs and excels in dialogue, translation, and reasoning. Open-source with weights and code.

Qwen2.5-Omni-3B. Developed by Alibaba, Qwen2.5-Omni-3B shares the same dual-path architecture as its larger counterpart but adopts a smaller backbone for efficient deployment. Despite its

reduced size, it retains competitive audio understanding and reasoning ability. Open-source with weights and code.

R1-AQA. Built on Qwen2-Audio-7B-Instruct and optimized with the GRPO reinforcement learning algorithm, R1-AQA is trained for audio question answering with only 38k post-training samples. Open-source with weights and code.

Audio-Reasoner. Built on Qwen2-Audio-Instruct and fine-tuned on CoTA, a large-scale chain-of-thought dataset, Audio-Reasoner adopts a four-stage structured reasoning template (Planning, Caption, Reasoning, Summary) to support deep audio reasoning. Open-source with weights and code.

Phi-4-Multimodal. Developed by Microsoft, Phi-4-Multimodal integrates text, vision, and speech/audio inputs into a single model via a Mixture-of-LoRAs architecture that routes each modality through dedicated adapters without interference. Open-source with weights and code.

Qwen2-Audio-Instruct. Developed by Alibaba, Qwen2-Audio-Instruct is an instruction-tuned audio-language model built on Qwen2-Audio. It supports two interaction modes, Voice Chat and Audio Analysis, which are automatically distinguished without any system-prompt switching. Open-source with weights and code.

GAMA. Developed by UMD and Adobe, GAMA employs an Audio Q-Former to convert AST features into audio tokens. Through instruction tuning on a synthetic reasoning dataset, it attains strong performance on reasoning-intensive audio benchmarks. Open-source with weights and code.

LTU. Developed by MIT-IBM, LTU combines a frozen Vicuna model with an audio encoder through LoRA adapters, and is trained on OpenAQA for non-speech audio question answering. It consumes continuous audio tokens to support open-ended reasoning. Open-source with weights and code.

F.2 Closed-source Models

Qwen3.5-Omni-Plus / Flash. Developed by Alibaba, both are multimodal models in the Qwen Omni series supporting text, image, audio, and video inputs. Qwen3.5-Omni-Plus is the high-performance variant, while Qwen3.5-Omni-Flash is the lightweight, low-latency variant with improved streaming capabilities over Qwen3-Omni. Both are accessible via API.

Gemini 3.1 Pro. Developed by Google, Gemini 3.1 Pro is a flagship multimodal model supporting text, image, audio, and video with low latency and long context. Accessible via API.

G Benchmarks Compared

MMAU Sakshi et al. (2025) is a large-scale multi-task audio understanding and reasoning benchmark, comprising 10,000 human-annotated multiple-choice questions that evenly cover the three major domains of speech, sound, and music. It spans 27 distinct skills, divided into information extraction and reasoning tasks, and its audio clips average roughly 10 seconds in length. Our experiments adopt widely-used test-mini split, the subset with publicly released answers.

MMAR Ma et al. (2025a) is a benchmark for deep audio reasoning, comprising 1,000 meticulously curated multiple-choice questions whose audio is collected from real-world internet videos and spans sound, music, speech, and their mixed modalities. It organizes tasks into a four-layer taxonomy including Signal, Perception, Semantic, and Cultural, with 16 fine-grained sub-categories, and annotates every question with a CoT rationale to characterize the reasoning process. Its questions generally require multi-step reasoning, and some demand graduate-level perceptual and domain-specific knowledge.

G.1 Comparison Results

Table 7 compares AQA accuracy on MMAU, MMAR, and MMALR.

Accuracy gaps. Among the 14 models, mean accuracy is 62.9% on MMAU, 51.8% on MMAR, and 36.0% on MMALR, demonstrating the challenge that MMALR poses to current LALMs.

Rank shifts. MMALR also yields a ranking that differs from those of prior benchmarks. Specifically, we analyze the ranking differences between MMALR and MMAU and MMAR. The results show that the Hamming distance is $11/14$ versus MMAU and $6/14$ versus MMAR, while the Cayley distance (minimum number of transpositions) is 8 and 4, respectively.

Table 7: AQA accuracy (%) on MMAU (test-mini), MMAR, and MMALR. Rank is the normal-condition rank among all 14 models (1 = best). **Bold** denotes the best score within each group.

	MMAU				MMAR				MMALR			
Model	Nor.	Noise	Silent	Rank	Nor.	Noise	Silent	Rank	Nor.	Noise	Silent	Rank
Open-source Models												
MiMo-Audio-Instruct	70.6	50.0	53.0	6	**60.6**	**38.9**	**41.1**	4	**38.5**	18.0	19.4	4
Audio Flamingo 3	**74.5**	**57.6**	**58.0**	4	53.2	37.1	39.1	5	37.6	**22.3**	**23.8**	5
Kimi-Audio-Instruct	54.9	38.3	38.4	10	47.0	25.3	26.1	9	32.8	16.7	17.8	8
Qwen2.5-Omni-7B	68.3	48.0	50.5	7	51.9	34.3	35.3	8	36.1	17.2	16.6	6
Qwen2.5-Omni-3B	71.0	47.2	48.5	5	52.8	33.8	35.3	6	33.3	16.2	16.4	7
R1-AQA	68.0	44.1	46.2	8	52.7	35.6	36.8	7	30.7	18.0	16.8	9
Audio-Reasoner	63.1	40.8	46.6	9	43.2	31.1	33.4	10	28.8	19.4	19.8	11
Phi-4-Multimodal	52.7	44.1	45.3	11	38.3	28.6	30.9	11	28.9	18.3	19.6	10
Qwen2-Audio-Instruct	50.4	31.9	32.2	12	33.5	22.0	23.1	12	27.1	16.4	16.9	12
GAMA	40.0	33.8	34.0	13	31.4	32.3	32.3	13	21.6	20.1	22.2	13
LTU	32.2	32.8	30.6	14	29.4	25.7	26.6	14	19.6	19.2	19.5	14
Closed-source Models												
Qwen3.5-Omni-Plus	**81.4**	**54.2**	**61.9**	1	78.7	**47.4**	**52.5**	2	57.2	**25.0**	**26.5**	2
Qwen3.5-Omni-Flash	78.1	48.3	54.2	2	73.0	43.2	48.0	3	46.4	20.0	21.1	3
Gemini 3.1 Pro	75.8	46.4	23.8	3	**79.2**	44.2	44.9	1	**65.0**	22.9	22.8	1

Within-model condition gaps. Table 8 reports the accuracy gap between normal and noise conditions, $\Delta = \mathrm{Acc}(\mathrm{normal}) - \mathrm{Acc}(\mathrm{noise})$, together with the corresponding normal–silent gap. Across the same 14 models, the mean audio gain is $+18.9$, $+17.5$, and $+16.7\%$ on MMAU, MMAR, and MMALR, respectively. This decreasing trend may be associated with the difficulty of leveraging audio evidence and the difficulty of overcoming textual priors. Closed-source models exhibit stronger audio utilization: Gemini 3.1 Pro reaches $+42.1\%$ on MMALR and $+35.0\%$ on MMAR, Qwen3.5-Omni-Plus reaches $+32.2\%$ on MMALR and $+31.3\%$ on MMAR, and Qwen3.5-Omni-Flash reaches $+29.8$, $+29.8$, and $+26.4\%$ on MMAU, MMAR, and MMALR. In contrast, early LALMs show limited audio gain: LTU is essentially flat on MMALR ($+0.4\%$), and GAMA reaches $+1.5\%$ on MMALR and -0.9% on MMAR. This pattern is consistent with the weaker audio analysis ability of early models.

TPE prior coverage. Prior coverage is the fraction of samples on which a model gives the same incorrect answer under both noise and silent conditions. Table 9 applies the same experimental setting to all three benchmarks. Across the 14 evaluated models, mean coverage is 30.2% on MMAU (range 15.6-40.7%), 38.6% on MMAR (23.4-54.8%), and 64.0% on MMALR (43.0–76.1%). This difference stems from the counter-prior design, which constructs more challenging textual-prior distractors, yielding more stable textual priors that enter the conditional ER/AR/OWT analysis and thereby making text-audio modality dependency measurable and decomposable.

H Additional Inference Details

Noise Input. In the noise condition, the original recording is replaced with Gaussian noise. Specifically, we use a scale of 0.02 for waveform inputs and 0.5 for filter-bank features.

Inference. For open-source models, all inference is conducted on NVIDIA A100 GPUs. Closed-source models are accessed through their respective APIs. For all models, we use consistent temperature settings, with temperature set to 0.0 for deterministic decoding where available.

Table 8: Within-model condition accuracy gaps (%).

	Δ normal−noise			Δ normal−silent		
Model	MMAU	MMAR	MMALR	MMAU	MMAR	MMALR
Open-source Models						
MiMo-Audio-Instruct	+20.6	+21.7	+20.5	+17.6	+19.5	+19.1
Audio Flamingo 3	+16.9	+16.1	+15.3	+16.5	+14.1	+13.8
Kimi-Audio-Instruct	+16.6	+21.7	+16.1	+16.5	+20.9	+15.0
Qwen2.5-Omni-7B	+20.3	+17.6	+18.9	+17.8	+16.6	+19.5
Qwen2.5-Omni-3B	+23.8	+19.0	+17.1	+22.5	+17.5	+16.9
R1-AQA	+23.9	+17.1	+12.7	+21.8	+15.9	+13.9
Audio-Reasoner	+22.3	+12.1	+9.4	+16.5	+9.8	+9.0
Phi-4-Multimodal	+8.6	+9.7	+10.6	+7.4	+7.4	+9.3
Qwen2-Audio-Instruct	+18.5	+11.5	+10.7	+18.2	+10.4	+10.2
GAMA	+6.2	−0.9	+1.5	+6.0	−0.9	−0.6
LTU	−0.6	+3.7	+0.4	+1.6	+2.8	+0.1
Closed-source Models						
Qwen3.5-Omni-Plus	+27.2	+31.3	+32.2	+19.5	+26.2	+30.7
Qwen3.5-Omni-Flash	+29.8	+29.8	+26.4	+23.9	+25.0	+25.3
Gemini 3.1 Pro	+29.4	+35.0	+42.1	+52.0	+34.3	+42.2
Mean	+18.9	+17.5	+16.7	+18.4	+15.7	+16.0

Table 9: Prior coverage on TPE sets. Coverage $= |E_m^{\mathrm{TPE}}|/N$.

Model	MMAU	MMAR	MMALR
Open-source Models			
MiMo-Audio-Instruct	27.5	40.9	70.9
Audio Flamingo 3	25.5	43.1	67.0
Kimi-Audio-Instruct	24.1	31.5	75.8
Qwen2.5-Omni-7B	35.0	49.6	70.2
Qwen2.5-Omni-3B	39.8	54.8	76.1
R1-AQA	33.2	44.5	61.5
Audio-Reasoner	19.3	23.4	43.0
Phi-4-Multimodal	35.0	47.4	62.7
Qwen2-Audio-Instruct	35.6	33.2	53.6
GAMA	40.7	44.5	58.6
LTU	33.9	40.7	74.5
Closed-source Models			
Qwen3.5-Omni-Plus	25.7	30.1	59.5
Qwen3.5-Omni-Flash	31.4	32.3	65.8
Gemini 3.1 Pro	15.6	23.7	56.6
Mean (14 models)	30.2	38.6	64.0

I Comparison with Related work

Hallucination. Hallucination typically refers to generated content that lacks support from the input audio, is inconsistent with the input audio, or has no basis in the input audio, and is primarily a factual or content-consistency error occurring at the model output level Cheng et al. (2026); Zhao et al. (2026). Our work focuses on a different phenomenon, which we call **text-prior dominance**: the tendency of a model to follow textual expectations while failing to make full use of acoustic evidence in its judgment. The distinction is that hallucination concerns whether the generated content is supported by the audio, whereas text-prior dominance concerns how a model decides between plausible textual preferences and genuine acoustic evidence. To our knowledge, MMALR is the first

to explicitly establish this controlled divergence within a multi-domain, multi-dimensional LALM evaluation framework, and to quantitatively observe the phenomenon of text-prior dominance.

Open-Ended AQA Open-ended AQA benchmarks assess a model's ability to formulate answers without predefined options. AIR-Bench Chat Yang et al. (2024) uses this format to evaluate generative audio understanding, while MMAU-Pro Kumar et al. (2026) includes open-ended questions alongside multiple-choice questions. Unlike these works, MMALR targets closed-ended AQA, where a model selects an answer from given options, and thus belongs to a different evaluation setting. A main motivation for open-ended AQA is to avoid the restricted answer space imposed by closed-ended options, but this format also introduces new evaluation challenges: answers may differ in wording, granularity, and specificity, so automatic evaluation often relies on additional semantic similarity judgments or LLM judges, increasing evaluation uncertainty. MMALR addresses a different question: in current LALM AQA evaluation, text-solvable samples can make evaluation unreliable, and textual priors may shape model predictions in AQA. To avoid the uncertainty of open-ended AQA in answer matching and semantic judgment, we conduct our study under closed-ended AQA. By keeping the question and options fixed across noise, silence, and normal-audio conditions, MMALR enables paired option-prediction analysis to quantify model reliance on textual priors.

Other studies on text-solvable samples in LALM evaluation. At the benchmark-construction level, ADQA-Bench He et al. (2026b) strengthens the audio dependency of samples through model-pool-based inference filtering and mainly evaluates models' listening or perceptual judgments. DEAF Xiong et al. (2026) constructs three types of acoustic-semantic conflicts in the speech domain, namely emotional prosody, background sounds, and speaker identity, and diagnoses model reliance on textual cues versus acoustic signals through progressive textual interference and multiple metrics. Unlike these works, MMALR is not confined to a specific audio domain or conflict type. Instead, it embeds counter-prior construction into multi-domain, multi-task AQA samples, jointly covering listening, understanding, and reasoning, and provides sample-level task categories, descriptions, and CoT annotations. At the evaluation-protocol level, Foo et al. (2026) propose text-prior and audio-reliance metrics and measures model reliance on acoustic signals by comparing full-audio and segmented-audio settings. MMALR's counter-prior construction is complementary to this evaluation-oriented perspective. Overall, the key distinction of MMALR is that it not only grounds correct answers in human-verified audio evidence at the benchmark-construction stage, but also introduces counter-prior distractors constructed from textual priors. On this basis, MMALR uses paired answer comparisons under uninformative- and normal-audio conditions to further analyze the roles of textual priors and real audio in model decision-making for AQA.

J Final Meta-Data Format

Each MMALR sample is stored in a structured JSON format containing the audio path, question, answer choices, correct answer, audio domain, task category and sub-category, language, source information, as well as the audio caption and the reasoning chain. An example of the final meta-data format is shown in Listing 4.

Listing 4: JSON annotation of MMALR.


```
{
  "id": "yt_v=KuE6lw4YnLM_00-00-00_00-00-13",
  "audio_path": "./audio/yt_v=KuE6lw4YnLM_00-00-00_00-00-13.wav",
  "question": "What caused the change in mom's attitude?",
  "choices": [
    "A. Something upset or stressed her",
    "B. A deliberate accent",
    "C. Someone familiar having an affair",
    "D. The child pronouncing it correctly"
  ],
  "answer": "B. A deliberate accent",
  "domain": "speech",
  "category": "Cultural Layer",
  "sub-category": "Cultural and Contextual Reasoning",
  "language": "English, Korean",
```

```
  "source": "yt",
  "url": "https://www.youtube.com/shorts/KuE6lw4YnLM",
  "timestamp": "00:00:00,00:00:13",
  "caption": "A daughter rushes to share gossip with her Korean mother
    who is watching TV: 'Mom! Mr. Lee's wife caught him cheating!' The
    mother, confused, asks twice: 'Who? Who did you say?' The daughter
    loudly repeats in standard American English pronunciation several
    times: 'His wife! Wife!' But the mother remains blank and does not
    understand. Finally, the daughter, frustrated, takes a deep breath
    and shouts in an exaggerated Korean English accent: 'Wipuuh (Wa-ee-pu
    )!' The mother instantly lights up, understands immediately, and
    exclaims excitedly.",
  "thinking": "The girl originally communicates in fluent English, but
    after her mother repeatedly expresses confusion, she suddenly
    switches to a typical Korean-accented English pronunciation. The
    mother's emotional shift upon hearing this specific accent indicates
    that the word exists in her vocabulary, but her auditory perception
    system can only match the Korean pronunciation. Therefore, the answer
     is: A deliberate accent."
}
```