Towards Reliable Large Audio Language Model

Ziyang Ma[1], Xiquan Li[1], Yakun Song[1], Wenxi Chen[1], Chenpeng Du[2], Jian Wu[2],
Yuanzhe Chen[2], Zhuo Chen[2], Yuping Wang[2], Yuxuan Wang[2], Xie Chen[1,3†]
[1] X-LANCE Lab, School of Computer Science,
MoE Key Lab of Artificial Intelligence, Shanghai Jiao Tong University
[2] ByteDance, [3]Shanghai Innovation Institute

Abstract

Recent advancements in large audio language models (LALMs) have demonstrated impressive results and promising prospects in universal understanding and reasoning across speech, music, and general sound. However, these models still lack the ability to recognize their knowledge boundaries and refuse to answer questions they don't know proactively. While there have been successful attempts to enhance the reliability of LLMs, reliable LALMs remain largely unexplored. In this paper, we systematically investigate various approaches towards reliable LALMs, including training-free methods such as multi-modal chain-of-thought (MCoT), and training-based methods such as supervised fine-tuning (SFT). Besides, we identify the limitations of previous evaluation metrics and propose a new metric, the Reliability Gain Index (RGI), to assess the effectiveness of different reliable methods. Our findings suggest that both training-free and training-based methods enhance the reliability of LALMs to different extents. Moreover, we find that awareness of reliability is a "meta ability", which can be transferred across different audio modalities, although significant structural and content differences exist among sound, music, and speech.

1 Introduction

Large audio language models (LALMs) have emerged as a promising approach to address the complex challenges of universal understanding and reasoning across diverse audio modalities, including speech (Wang et al., 2023a; Hu et al., 2024; Deng et al., 2024a), music (Deng et al., 2024b; Liu et al., 2024b), and general sound (Gong et al., 2024; Kong et al., 2024). LALMs leverage the power of large-scale pre-trained encoders and LLMs to capture intricate acoustic features and semantic representations across various scenarios (Gong et al., 2023; Ghosh et al., 2024; Tang et al., 2024; Chu et al., 2023, 2024), demonstrating potential to handle a wide range of unsolved tasks. Despite their impressive performance, LALMs still face a significant limitation: they lack the ability to recognize when they do not know the answer to a question, as shown in Figure 1. The left part depicts an unreliable LALM, where the model provides either correct or incorrect answers for audio inputs (sound, music, or speech) without rejecting awareness. The right part shows a reliable LALM, where the model refuses to answer questions that exceed its knowledge boundary. This reliability helps prevent models from offering incorrect or overly confident responses when faced with uncertainty.

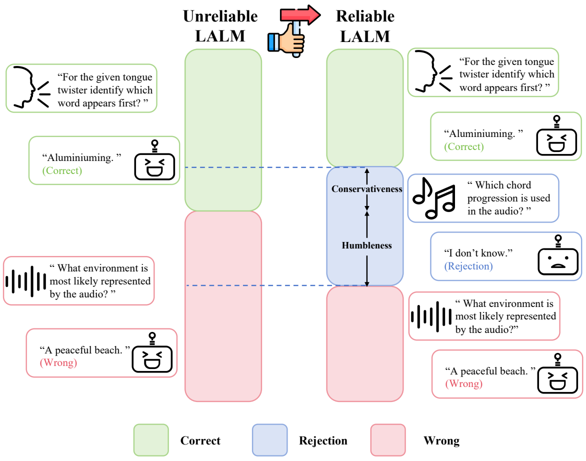

Unreliable LALM
Reliable LALM
"For the given tongue twister identify which word appears first? "
"Aluminiuming. " (Correct)
" What environment is most likely represented by the audio? "
"A peaceful beach. " (Wrong)
"For the given tongue twister identify which word appears first? "
"Aluminiuming. " (Correct)
Conservativeness
Humbleness
" Which chord progression is used in the audio? "
"I don't know." (Rejection)
" What environment is most likely represented by the audio?"
"A peaceful beach. " (Wrong)
🟩 Correct
🟦 Rejection
🟥 Wrong

Figure 1: Illustration of Unreliable vs. Reliable LALMs.

While there have been successful efforts to enhance the reliability of language models in text-based models (Yang et al., 2024d; Cheng et al., 2024a; Yona et al., 2024; Zhang et al., 2024; Xu et al., 2024a), reliable LALMs remain largely unexplored. Building reliability in LALMs is crucial for applications where the model's confidence is essential, especially in real-world scenarios such as healthcare (Jia et al., 2024), autonomous driving (Xu et al., 2024c), and interactive agents (Ma et al., 2025a). A reliable LALM should not only

[†]Corresponding author

Findings of the Association for Computational Linguistics: ACL 2025, pages 1000–1014
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

provide accurate answers but also have the ability to refuse to answer when it is unsure, offering a more responsible interaction model.

In this paper, we present a comprehensive investigation into methods for enhancing the reliability of LALMs. We explore both training-free and training-based approaches that can effectively improve LALM's ability to identify its knowledge boundaries and reject incorrect answers. Specifically, we investigate how these methods influence accuracy, truthfulness, and reliability in a variety of audio modalities. Furthermore, we introduce a novel quantitative metric, the Reliability Gain Index (RGI), to assess the effectiveness of different reliability-enhanced techniques. The goal is for the model to avoid being overly conservative (from what it knows), while promoting humility (from what it doesn't konw) by rejecting answering, as shown in Figure 1. Our experiments demonstrate that the ability to say "I don't know" (IDK) is a "meta ability" of LALMs, which means this ability can be trained on one modality and transferred to other audio modalities, even in the presence of significant structural and content differences among sound, music, and speech. Our contribution can be summarized in the following points:

1. To the best of our knowledge, we are the first to investigate the reliability of LALM. Both training-free and training-based methods are conducted to verify LALM's reliability.

2. We identify the limitations of previous evaluation metrics and propose a new metric, the Reliability Gain Index (RGI), to measure the effectiveness of different reliable methods.

3. With our new metric under the cross-modal setting, we find that the awareness of reliability is a "meta ability" and can be transferred to other modalities in the context of large audio language modeling.

2 Related Work

2.1 Large Audio Language Model

Large audio language models (LALMs), as a rising part of multimodal large language models (MLLMs), aim to leverage the capabilities of LLMs to achieve advanced audio understanding and reasoning abilities. However, the inherent differences in acoustic structures and content across speech, music, and general sound make universal audio processing challenging, primarily due to the domain conflict (Wang et al., 2023b) and the catastrophic forgetting (Tang et al., 2024) problem. Balancing the performance across these different modalities remains difficult.

Some works focus on modeling individual modalities with LLMs among speech, music, and general sound. In speech language models (SLMs), tasks such as LLM-based speech recognition (ASR) (Li et al., 2023b; Wu et al., 2023a; Ma et al., 2024; Yu et al., 2024; Yang et al., 2024b,a; Geng et al., 2024; Ma et al., 2025b), speaker diarization (SD) (Shi et al., 2024; Meng et al., 2024), and speech emotion recognition (SER) (Xu et al., 2024b; Lin et al., 2024; Cheng et al., 2024b; Kang et al., 2024) are critical, which involve either acoustic features, semantic features, or both, specific to speech. Furthermore, some works tackle various tasks with all-in-one modeling within the speech domain (Wang et al., 2023a; Hu et al., 2024; Deng et al., 2024a). In the realm of LLM-based music understanding, some approaches focus on solving problems related to signal-based music processing (Deng et al., 2024b; Liu et al., 2024b), while others target symbolic music understanding (Yuan et al., 2024; Qu et al., 2024). Within general sound understanding, LLM-based automated audio captioning (AAC) (Wu et al., 2023b; Chen et al., 2025; Li et al., 2025; Liu et al., 2024a) and subsequent audio question answering (AQA) task (Gong et al., 2024; Kong et al., 2024) have received significant attention due to their potential to advance the field. Further research has explored methods that can handle two (Gong et al., 2023; Ghosh et al., 2024) or three (Tang et al., 2024; Chu et al., 2023, 2024) modalities simultaneously by scaling data and model parameters, as well as innovating in the design of data pipelines (Lu et al., 2024a,b) and model architectures (Bhati et al., 2024).

The recently developed open-source model, Qwen2-Audio (Chu et al., 2024), demonstrates strong general-purpose audio understanding and reasoning capabilities across various benchmarks (Yang et al., 2024c; Li et al., 2024; Sakshi et al., 2024) through a combination of pre-training, supervised fine-tuning (SFT), and reinforcement learning with human feedback (RLHF). Despite these advances, even Qwen2-Audio still lacks awareness of reliability in our experiments. This highlights the need for further exploration into the reliable LALM, an area that remains an open challenge in the field.

2.2 Evaluation of Reliable Generation

Evaluating the reliability of LLMs is a new topic that lacks a unified standard. In general, the goal is for LLMs to be able to refuse answering when they are unable to derive an answer. Several approaches have been proposed to assess the reliability of LLMs. Yang et al. (2024d) introduced the use of Prudence Score, Over-Conservativeness Score, and Honesty Score to evaluate model performance. Cheng et al. (2024a) were the first to introduce the concept of the IDK dataset and employed Knowledge Quadrants to visualize a model's knowledge coverage and then defined the Truthfulness score. Yona et al. (2024) defined the Mean Faithful Generation (MFG) metric, which quantifies the expected faithfulness of a single model output by comparing it against a ground truth. Additionally, Zhang et al. (2024) proposed the Average Precision (AP) score measuring the model's precision in identifying and ranking relevant predictions based on its knowledge. A more recent and popular method for evaluating reliability involves weighting accuracy and truthfulness to obtain a final reliability score introduced by Xu et al. (2024a). This method provides a more holistic measure of a model's overall reliability by balancing both the model's correct responses and its ability to reject uncertain answers.

3 Reliable LALM

The core purpose of reliable LALM is to refuse to answer when the model doesn't know the answer given the input audio and instruction. We explored two distinct approaches to enhance the model's reliability: training-free and training-based methods. The training-free method aims to activate the inherent capabilities of the model through prompting or agent, thereby achieving reliability without requiring additional training. In contrast, the training-based method seeks to enhance the model's reliability by post-training, thereby explicitly advertising reliability through the training process.

3.1 Training-free Method

Three training-free methods for enhancing model reliability are employed: IDK Prompting, MCoT Prompting, and Task Agent. These methods leverage the model's inherent instruction-following capabilities to improve its reliability without additional training.

IDK Prompting. I Don't Know (IDK) Prompting is a method designed to enhance model reliability by adding a supplementary prompt after the input question. This prompt encourages the model to acknowledge uncertainty in cases where it lacks sufficient information to provide a confident answer. By incorporating this strategy, the model is prompted to explicitly state "I don't know" when necessary. The specific prompt used in this method is outlined in Appendix C.1.

MCoT Prompting. Multi-modal Chain-of-Thought (MCoT) Prompting (Lu et al., 2022; Zhang et al., 2023) encourages the LALM to reason step by step, with the intention of refining its analysis of the given problem and producing more reliable results. This approach leads the model to break down complex tasks into smaller, manageable components, which are processed sequentially. By prompting the model to articulate its thought process, MCoT Prompting improves its ability to reason logically and arrive at reliable conclusions. The specific prompt for MCoT is provided in Appendix C.1.

Task Agent. Audio data differs from textual data significantly, presenting substantial differences in the content and structure among speech, sound, and music, despite all being categorized as "audio". We propose to use a task agent, which is designed to support multi-step reasoning with tool-using ability during the inference process. This approach incorporates a sequence of steps for reasoning that includes identifying the type of audio, analyzing its content, and producing a final prediction. The steps are as follows:

1. **Identify the type of audio.** The model is first required to categorize the audio input. The possible types include sound, music, and speech.

2. **Generate content based on audio type.** Depending on the identified audio type, the model generates the corresponding content. For speech, the model outputs the automatic speech recognition (ASR) result. For sound or music, the model generates an automatic audio caption (AAC) or music caption (MC).

3. **Output generation.** After completing the previous steps, the model combines the audio, question, and generated content, and then inputs them into the LALM to obtain the final answer.

This multi-step reasoning process ensures that the model can make context-aware decisions, further enhancing its reliability. We provide the detailed prompt processing in Appendix C.1.

3.2 Training-based Method

Following previous works (Yang et al., 2024d; Cheng et al., 2024a; Zhang et al., 2024), we adopt a similar approach but apply it to the multi-modal setting. The training-based method involves two key steps: construction of a model-specific IDK dataset and post-training of the model.

Construction of the IDK Dataset. Given that different models possess varying knowledge quadrants, it is essential to construct a model-specific IDK dataset for each model. For each data point, N possible answers are sampled. If the model provides the correct answer at least K times (where $K \leq N$), we assume the model's knowledge adequately covers the question, and the original answer is retained as the ground truth. Conversely, if the model fails to answer correctly enough times, the answer is labeled as IDK. The threshold parameter, denoted as $K@N$, plays a critical role in defining the IDK dataset. $0@N$ means no IDK data is generated, where all answers retain their original labels. While $N@N$ means the model must answer all N times correctly to retain the original label. By adjusting this threshold, different knowledge levels of the model-specific IDK dataset can be generated.

Post-training with the IDK Dataset. Once the IDK dataset is constructed, we perform supervised fine-tuning (SFT) to align the model's reliability. Specifically, we utilize the IDK dataset to fine-tune the model, guiding it to better handle uncertainty and enhance its reliability. During this process, the model learns to recognize when it should confidently provide an answer and when it should appropriately output IDK.

4 Evaluation

We first introduce a basic evaluation metric for truthfulness proposed by Cheng et al. (2024a) and further, reliability proposed by Xu et al. (2024a). We then propose our new metric, Reliability Gain Index (RGI), to measure the effectiveness of different reliable methods.

4.1 Reliability Evaluation

To evaluate the reliability of the given LALM, we consider using the weighted overall reliability that balances the model's helpfulness and truthfulness. Let N denote the total number of queries tested, which can be expressed as:

$$N = N_c + N_r + N_w, \tag{1}$$

where N_c, N_r, N_w donate the numbers of correct, rejected, and wrong answers. Accuracy (Acc) measures the proportion of correct answers relative to the total number of queries. It is calculated as:

$$Acc = \frac{N_c}{N}. \tag{2}$$

Truthfulness (Tru) quantifies how truthful the model's answers are when it does not reject a query. It is defined as the proportion of not wrong among queries, which is given by:

$$Tru = 1 - \frac{N_w}{N}. \tag{3}$$

Rejection Rate (Rej) measures the fraction of queries for which the model chooses to refuse to answer. It is calculated as:

$$Rej = \frac{N_r}{N} \tag{4}$$

The Rejection Rate reflects the model's willingness to reject uncertain or out-of-scope queries, thereby avoiding providing potentially unreliable answers. Reliability (Rel) combines the model's accuracy and truthfulness into a single measure. It accounts for both the model's performance on correct answers and its ability to reject uncertain responses. The Reliability score is given by:

$$Rel = Rej \cdot Acc + (1 - Rej) \cdot Tru, \tag{5}$$

where the Reliability is weighted by the rejection rate. The Rejection Rate represents the degrees of sensitivity towards errors.

4.2 Reliability Gain

While the previously defined metrics such as Accuracy, Truthfulness, Rejection Rate, and Reliability are useful for evaluating the absolute capability of a model to express IDK, they are less effective in measuring the relative effectiveness of different reliable methods. Specifically, these metrics fail to reveal how well a method balances two crucial aspects of reliability: conservativeness (the tendency to reject correct answers) and humbleness (the tendency to reject incorrect answers).

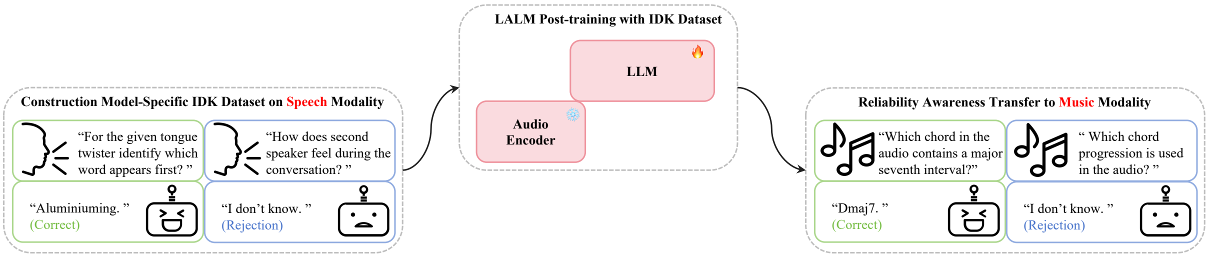

LALM Post-training with IDK Dataset
LLM
Audio Encoder
Construction Model-Specific IDK Dataset on Speech Modality
“For the given tongue twister identify which word appears first? ”
“How does second speaker feel during the conversation? ”
“Aluminiuming. ” (Correct)
“I don’t know. ” (Rejection)
Reliability Awareness Transfer to Music Modality
“Which chord in the audio contains a major seventh interval?”
“ Which chord progression is used in the audio? ”
“Dmaj7. ” (Correct)
“I don’t know. ” (Rejection)

Figure 2: Illustration of Reliability Awareness Transferability in LALM.

To capture the effectiveness of reliable methods more precisely, we introduce the Reliability Gain Index (RGI). Specifically, the RGI evaluates the relative gains in the model’s rejection capabilities, distinguishing between increases in relative conservativeness and relative humbleness. Let N_c and N_w denote the numbers of correct and incorrect answers, respectively, in the original unreliable model:

$$N = N_c + N_w. \tag{6}$$

After applying a reliable method (e.g., prompting or supervised fine-tuning), the numbers of correct and wrong answers may be redistributed into different categories:

$$\begin{aligned} N_c &= N_{cc} + N_{cr} + N_{cw}, \\ N_w &= N_{wc} + N_{wr} + N_{ww}, \end{aligned} \tag{7}$$

where N_{cc} refers to correct answers that remain correct, N_{cr} refers to correct answers that are rejected, and N_{cw} refers to correct answers that are reclassified as wrong. The same marks are also applied to N_w.

To quantify how much the model has become more conservative, we define the relative conservativeness increase as:

$$\Delta_{Con} = \frac{N_c - N_{cc}}{N_c}. \tag{8}$$

This metric captures the proportion of correct answers that were rejected after applying the reliable method. A higher value of Δ_{Con} indicates that the model has become more conservative in its response, rejecting more previously correct answers. To measure the increase in the model’s humbleness, we define the relative humbleness increase as:

$$\Delta_{Hum} = \frac{N_w - N_{ww}}{N_w} \tag{9}$$

This metric reflects the proportion of wrong answers that were correctly rejected (i.e., converted into IDK). A higher value of Δ_{Hum} indicates that the model has become more humble by rejecting answers beyond its knowledge or capacity. Finally, we introduce the Reliability Gain Index (RGI) to combine the changes in conservativeness and humbleness. The RGI is defined as:

$$RGI = \log(\frac{\Delta_{Hum}}{\Delta_{Con}}) \tag{10}$$

This metric provides a measure of the model’s improvement in reliability. A higher RGI value reflects that the model has become more humble while avoiding excessive conservatism, as it demonstrates a favorable index between increasing the rejection of wrong answers and minimizing the rejection of correct ones. For a concrete example illustrating why traditional metrics may inadequately capture the effectiveness of reliability methods, and a formal derivation of the conditions under which these metrics may be invalid, refer to Appendix E.

4.3 Cross-modal Reliability Awareness

An intriguing question is whether the awareness of reliability can transfer across different audio modalities, as shown in Figure 2. Using the proposed RGI metric, we can easily answer this question. Specifically, if we train the model on one modality and test it on another, an $RGI > 0$ indicates that the model has successfully learned to reject more questions it does not know, even when tested on a different audio modality. Such transferability is crucial for building robust models that can operate reliably across various domains and even various modalities.

5 Experiments

5.1 Setup

We use Qwen2-Audio-7B-Instruct (Chu et al., 2024) as the baseline model in our experiments. This model has demonstrated strong performance across various benchmarks, including AIR-Bench (Yang et al., 2024c), OmniBench (Li

Table 1: *Accuracy* (Acc%↑), *Truthfulness* (Tru%↑), and *Reliability* (Rel%↑) performance comparison of Qwen2-Audio-7B-Instruct baselines, training-free methods, and training-based methods on the MMAU benchmark across sound, speech, and music modalities. The result for LoRA Fine-tuning is computed by cross-validation across three modalities. The best-performing items are highlighted in **bold**, and the second-best items are underlined. We also show random guess, most frequent choice, and human evaluation results from the original MMAU paper for reference.

Methods	**Post Training**	**Sound**			**Music**			**Speech**			**Total**		
		Acc%	Tru%	Rel%	Acc%	Tru%	Rel%	Acc%	Tru%	Rel%	Acc%	Tru%	Rel%
Baseline													
MMAU (Unnormalized)	-	54.95	54.95	54.95	50.98	50.98	50.98	42.04	42.04	42.04	49.20	49.20	49.20
Ours (Normalized)	-	60.96	60.96	60.96	**55.09**	55.09	55.09	**50.75**	50.75	50.75	**55.60**	55.60	55.60
Reliable LALM													
IDK Prompting	✗	58.26	**76.28**	**73.03**	54.19	66.77	65.19	43.84	58.26	56.18	52.10	67.10	64.85
MCoT Prompting	✗	57.96	68.17	67.13	51.50	**71.56**	67.53	44.74	60.06	57.71	51.40	66.60	64.29
Task Agent	✗	58.56	72.67	70.68	53.29	**71.56**	**68.22**	46.25	59.76	57.93	52.70	68.00	65.66
LoRA Fine-tuning	✓	**61.71**	71.77	70.71	51.35	70.66	66.43	47.90	**61.86**	**59.91**	53.65	**68.10**	**65.68**
Reference													
Random Guess	-	26.72	26.72	26.72	24.55	24.55	24.55	26.72	26.72	26.72	26.00	26.00	26.00
Most Frequent Choice	-	27.02	27.02	27.02	20.35	20.35	20.35	29.12	29.12	29.12	25.50	25.50	25.50
Human	-	86.31	86.31	86.31	78.22	78.22	78.22	82.17	82.17	82.17	82.23	82.23	82.23

et al., 2024), and MMAU (Sakshi et al., 2024), positioning it as one of the most powerful open-source LALMs available. We also test the performance of other LALMs, whose detailed introduction can be found in Appendix B.1, and respective performance are shown in Appendix B.2.

Our experiments are conducted on the MMAU dataset, which consists of human-annotated natural language questions and answers covering three domains: speech, environmental sounds, and music. Details of the dataset can be found in Appendix A.

For the construction of the IDK dataset, we employ a 5@5 threshold, following Cheng et al. (2024a)'s setting. Specifically, for each given question, we perform 5 rounds of inference with the LALM model. If the model answers correctly in all 5 rounds, we consider the model to have sufficient knowledge of the question, and the original answer is retained as the ground truth. However, if the model answers incorrectly in any of the 5 rounds, the answer is labeled as IDK.

For the training-free methods, we employ both single-step and multi-step reasoning introduced in Section 3.1. The specific prompts used in these approaches are provided in Appendix C.1. For the training-based methods, since the Qwen-Audio series does not provide fine-tuning code, we implement our fine-tuning process based on DeepSpeed [1], using Low-Rank Adaptation (LoRA) (Hu et al., 2021) with Parameter Efficient Fine-Tuning (PEFT) library [2]. For all three modalities, we perform SFT on the model-specific IDK dataset for 1 epoch. Detailed training hyper-parameters for each modality are provided in Appendix D.

5.2 Reliable Methods Analysis

Table 1 presents the Accuracy, Truthfulness, and Reliability metrics for different reliable methods. For the baseline on the MMAU dataset, the answers are extracted with rule-based methods, which results in inaccurate judgments of the model's output. To address this, we further utilized the *GPT-4o-mini* API to regularize the answers. The prompt template for answer normalization is shown in Appendix C.2. From the table, it is evident that for the different reliable methods, the model's Accuracy generally decreases compared to the baseline. However, Truthfulness improves consistently, and the Reliability metric, a weighted balance between Accuracy and Truthfulness, also shows an increase, indicating an overall improvement in the model's reliability. Notably, the training-free methods achieve high Truthfulness, but their impact on Accuracy is relatively large. This suggests that the training-free methods tend to make the model more conservative or humble, resulting in a higher rejection rate, negatively affecting the helpfulness. In contrast, the training-based methods demonstrate a better trade-

[1] https://github.com/microsoft/DeepSpeed

[2] https://github.com/huggingface/peft

Table 2: *Relative Conservativeness Increase* ($\Delta_{Con}\%\downarrow$), *Relative Humbleness Increase* ($\Delta_{Hum}\%\uparrow$) and *Reliability Gain Index* (RGI↑) performance of Qwen2-Audio-7B-Instruct with different reliable methods on the MMAU benchmark across sound, speech, and music modalities. The result for LoRA Fine-tuning is computed by cross-validation across three modalities. The best-performing items are highlighted in **bold**, and the second-best items are underlined.

Methods	**Post Training**	**Sound**			**Music**			**Speech**			**Total**		
		$\Delta_{Con}\%$	$\Delta_{Hum}\%$	RGI	$\Delta_{Con}\%$	$\Delta_{Hum}\%$	RGI	$\Delta_{Con}\%$	$\Delta_{Hum}\%$	RGI	$\Delta_{Con}\%$	$\Delta_{Hum}\%$	RGI
IDK Prompting	✗	10.81	**20.12**	0.27	12.87	20.36	0.20	15.61	16.52	0.02	13.10	**19.00**	0.16
MCoT Prompting	✗	11.71	14.41	0.09	11.68	20.96	0.25	13.21	16.22	0.09	12.20	17.20	0.15
Task Agent	✗	9.61	16.52	0.24	**11.38**	20.96	**0.27**	**9.61**	14.11	0.17	**10.20**	17.20	0.23
LoRA Fine-tuning	✓	**6.91**	15.62	**0.36**	12.73	**21.11**	0.23	11.56	**18.17**	**0.19**	10.40	18.30	**0.26**

off between Accuracy and Truthfulness, as they manage to improve Truthfulness while minimizing the negative impact on Accuracy. Consequently, the Reliability of the model is improved in a more balanced manner.

Table 2 presents the Relative Conservativeness Increase, Relative Humbleness Increase, and Reliability Gain Index for different reliable methods. From Table 1, increasing both conservativeness and humbleness will result in an improvement in model reliability. However, a greater increase in humbleness than in conservativeness is key to achieving an effective reliable method. From Table 2 we observe that, regardless of training-free or training-based methods, the RGI is greater than 0 across all modalities, indicating that different reliable methods are generally effective. When analyzing the results by modality, we find that the RGI is higher for the sound and music, while it is relatively lower for the speech. This suggests that the model is more confident about what it knows and does not know in the sound and music. The use of Task Agent (i.e., ASR results) or SFT helps mitigate this issue in speech, improving the model's reliability. From all these results, it is evident that training-based methods strike a better trade-off between conservativeness and humbleness, thereby achieving a superior RGI compared to the training-free methods.

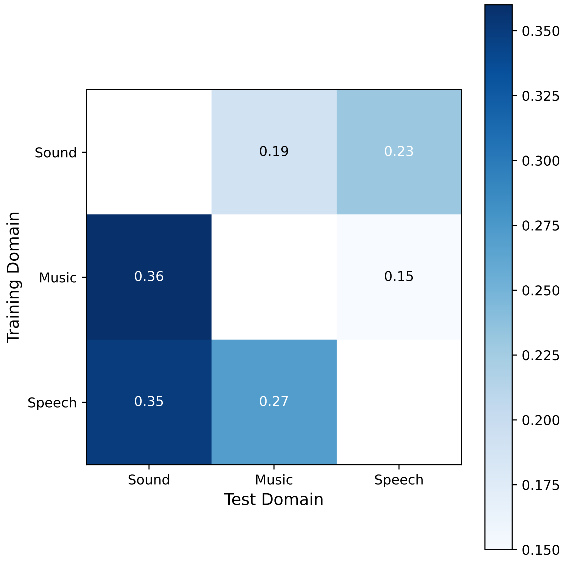


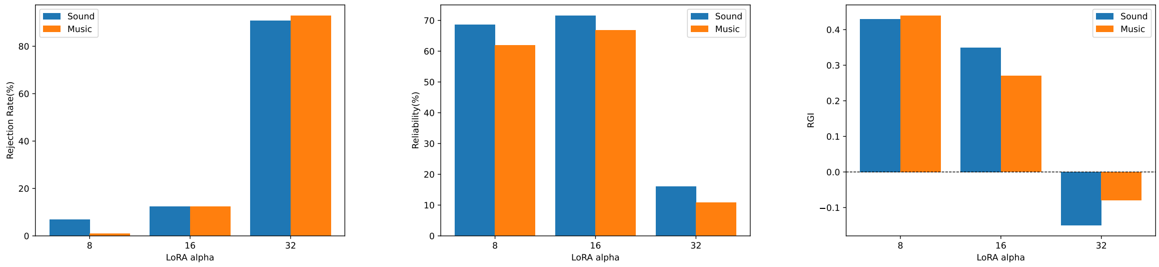
Figure 3: Heatmap for cross-modal SFT results. The figure shows the RGI performance of reliable models trained on one modality and tested on another.

5.3 Cross-modal Analysis

Figure 3 presents the results of cross-model SFT. Despite the significant structural and content differences among the sound, music, and speech modalities in audio processing, the heatmap reveals that all RGI values are greater than 0. This indicates that the LALM's ability to express "I don't know" is a "meta ability", which can be learned in one modality and transferred to others. Notably, training on one modality and testing on another often results in a high RGI when tested on the sound modality, suggesting that the model's knowledge is more distinctly separable on sound compared to the other modalities. This implies that sound tasks provide clearer boundaries for what the model knows and does not know. Detailed cross-modal testing results can be found in Appendix F.1.

5.4 Ability Study

Figure 5 illustrates the percentage for constructing IDK dataset using different $K@N$ thresholds. In our experiments, N is set to 5. As the $K@5$ threshold increases, the requirement for model's certainty becomes stricter, resulting in a higher percentage of IDK data. However, the change in IDK percentage from $1@5 = 50.2\%$ to $5@5 = 63.5\%$ is relatively small, compared to the text modality (Cheng et al., 2024a). This suggests that although the capability of LALMs still requires improvement compared to the text modality, their response stability is rel-



(a) (b) (c)

Figure 4: Rejection Rate(%), Reliability(%), and Reliability Gain Index (RGI) performance for different LoRA alpha weights trained on speech modality.

atively high, indicating a strong foundation to develop reliable methods.

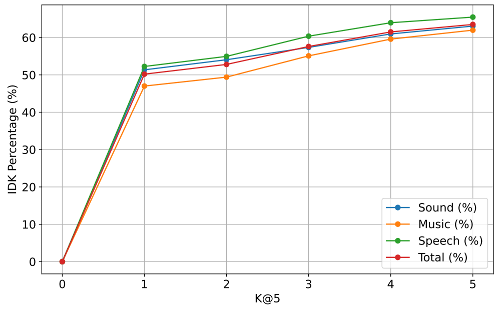


Figure 5: IDK percentage for constructing IDK dataset with different $K@5$ threshold.

The selection of LoRA weights is crucial for balancing between helpfulness and truthfulness. Figure 4 illustrates the impact of various LoRA alpha weights on Rejection Rate (%), Reliability (%), and Reliability Gain Index (RGI). The Rejection Rate for the main results is provided in Appendix F.2. The model undergoes SFT for reliability on the speech modality and is tested on the sound and music modalities. As shown in Figure 4(a), the Rejection Rate increases with the LoRA alpha weight grows, indicating that a smaller LoRA alpha weight prevents the model from learning to reject unknown answers, while a larger LoRA alpha weight leads to over-conservatism. In Figure 4(b), the Reliability metric initially increases with the LoRA alpha weight but eventually decreases, demonstrating a non-monotonic relationship. Figure 4(c) shows that the RGI value decreases as the LoRA alpha weight grows, reaching a point when the training becomes ineffective (when RGI < 0). Interestingly, very small LoRA alpha weights can also achieve high RGI values, suggesting that the awareness of reliability is relatively easy to acquire and transfer across modalities.

6 Conclusion & Future Work

In this work, we have systematically investigated the reliability of large audio language models (LALMs), introducing both training-free and training-based methods to reject questions the LALM cannot answer. We propose the novel Reliability Gain Index (RGI) metric, which quantifies the effectiveness of different reliable methods in improving model reliability. We have demonstrated that awareness of reliability is a “meta ability” of the model, and this awareness can be transferred across various audio modalities, including speech, sound, and music, even when these modalities differ significantly in structure and content. Our findings contribute to the ongoing efforts to build more reliable LALMs and provide a foundation for future work in this direction. While our study has investigated the transferability of reliability awareness across different audio modalities, future work will explore the possibility of transferring this capability between even more disparate modalities, such as the speech and video modalities, within the context of an Omni Language Model (OLM).

Acknowledge

This work was supported by the National Natural Science Foundation of China (No. 62206171 and No. U23B2018), Shanghai Municipal Science and Technology Major Project under Grant 2021SHZDZX0102 and Yangtze River Delta Science and Technology Innovation Community Joint Research Project (2024CSJGG01100).

Limitation

While this work has made significant strides in investigating the reliability of LALMs by focusing on their ability to reject questions with “I don’t know”, it primarily addresses the basic aspect of model reliability. Specifically, our study does not explore the potential for the model to provide more detailed justifications for its refusal. A promising direction for future research is to enable models to actively ask for additional information in an interactive manner when they are unsure, and to provide more reliable answers based on a deeper understanding of the user’s query. This would not only enhance the model’s reliability awareness but also make it more context-aware and capable of engaging in dynamic interactions with users, ultimately leading to more intelligent and trustworthy responses.

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Saurabhchand Bhati, Yuan Gong, Leonid Karlinsky, Hilde Kuehne, Rogerio Feris, and James Glass. 2024. State-space large audio language models. *arXiv preprint arXiv:2411.15685*.

Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, and Furu Wei. 2023. BEATs: Audio pre-training with acoustic tokenizers. *Proc. ICML*.

Wenxi Chen, Ziyang Ma, Xiquan Li, Xuenan Xu, Yuzhe Liang, Zhisheng Zheng, Kai Yu, and Xie Chen. 2025. SLAM-AAC: Enhancing audio captioning with paraphrasing augmentation and CLAP-Refine through llms. *Proc. ICASSP*.

Qinyuan Cheng, Tianxiang Sun, Xiangyang Liu, Wenwei Zhang, Zhangyue Yin, Shimin Li, Linyang Li, Zhengfu He, Kai Chen, and Xipeng Qiu. 2024a. Can ai assistants know what they don’t know? *Proc. ICML*.

Zebang Cheng, Zhi-Qi Cheng, Jun-Yan He, Jingdong Sun, Kai Wang, Yuxiang Lin, Zheng Lian, Xiaojiang Peng, and Alexander Hauptmann. 2024b. Emotion-LLaMA: Multimodal emotion recognition and reasoning with instruction tuning. *Proc. NeurIPS*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*.

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

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. 2023. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*.

Keqi Deng, Guangzhi Sun, and Philip C Woodland. 2024a. Wav2prompt: End-to-end speech prompt generation and tuning for llm in zero and few-shot learning. *arXiv preprint arXiv:2406.00522*.

Zihao Deng, Yinghao Ma, Yudong Liu, Rongchen Guo, Ge Zhang, Wenhu Chen, Wenhao Huang, and Emmanouil Benetos. 2024b. Musilingo: Bridging music and text with pre-trained language models for music captioning and query response. *Proc. NAACL*.

Xuelong Geng, Tianyi Xu, Kun Wei, Bingshen Mu, Hongfei Xue, He Wang, Yangze Li, Pengcheng Guo, Yuhang Dai, Longhao Li, et al. 2024. Unveiling the potential of llm-based asr on chinese open-source datasets. *arXiv preprint arXiv:2405.02132*.

Sreyan Ghosh, Sonal Kumar, Ashish Seth, Chandra Kiran Reddy Evuru, Utkarsh Tyagi, S Sakshi, Oriol Nieto, Ramani Duraiswami, and Dinesh Manocha. 2024. GAMA: A large audio-language model with advanced audio understanding and complex reasoning abilities. *Proc. EMNLP*.

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

Yuan Gong, Hongyin Luo, Alexander H Liu, Leonid Karlinsky, and James Glass. 2024. Listen, think, and understand. *Proc. ICLR*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Shujie Hu, Long Zhou, Shujie Liu, Sanyuan Chen, Lingwei Meng, Hongkun Hao, Jing Pan, Xunying Liu, Jinyu Li, Sunit Sivasankaran, et al. 2024. Wavllm: Towards robust and adaptive speech large language model. *Proc. EMNLP*.

Shuyue Jia, Subhrangshu Bit, Edward Searls, Lindsey Claus, Pengrui Fan, Varuna H Jasodanand, Meagan V Lauber, Divya Veerapaneni, William M Wang, Rhoda Au, et al. 2024. MedPodGPT: A multilingual audio-augmented large language model for medical research and education. *medRxiv*.

Wonjune Kang, Junteng Jia, Chunyang Wu, Wei Zhou, Egor Lakomkin, Yashesh Gaur, Leda Sari, Suyoun Kim, Ke Li, Jay Mahadeokar, et al. 2024. Frozen large language models can perceive paralinguistic aspects of speech. *arXiv preprint arXiv:2410.01162*.

Zhifeng Kong, Arushi Goel, Rohan Badlani, Wei Ping, Rafael Valle, and Bryan Catanzaro. 2024. Audio flamingo: A novel audio language model with few-shot learning and dialogue abilities. *Proc. ICML*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *Proc. ICML*.

Xiquan Li, Wenxi Chen, Ziyang Ma, Xuenan Xu, Yuzhe Liang, Zhisheng Zheng, Qiuqiang Kong, and Xie Chen. 2025. DRCap: Decoding CLAP latents with retrieval-augmented generation for zero-shot audio captioning. *Proc. ICASSP*.

Yizhi Li, Ge Zhang, Yinghao Ma, Ruibin Yuan, Kang Zhu, Hangyu Guo, Yiming Liang, Jiaheng Liu, Zekun Wang, Jian Yang, et al. 2024. Omnibench: Towards the future of universal omni-language models. *arXiv preprint arXiv:2409.15272*.

Yuang Li, Yu Wu, Jinyu Li, and Shujie Liu. 2023b. Prompting large language models for zero-shot domain adaptation in speech recognition. *Proc. ASRU*.

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

Jizhong Liu, Gang Li, Junbo Zhang, Chenyu Liu, Heinrich Dinkel, Yongqing Wang, Zhiyong Yan, Yujun Wang, and Bin Wang. 2024a. Leveraging ced encoder and large language models for automated audio captioning. *Proc. DCASE Challenge*.

Shansong Liu, Atin Sakkeer Hussain, Chenshuo Sun, and Ying Shan. 2024b. Music understanding llama: Advancing text-to-music generation with question answering and captioning. *Proc. ICASSP*.

Ke-Han Lu, Zhehuai Chen, Szu-Wei Fu, He Huang, Boris Ginsburg, Yu-Chiang Frank Wang, and Hung-yi Lee. 2024a. Desta: Enhancing speech language models through descriptive speech-text alignment. *Proc. Interspeech*.

Ke-Han Lu, Zhehuai Chen, Szu-Wei Fu, Chao-Han Huck Yang, Jagadeesh Balam, Boris Ginsburg, Yu-Chiang Frank Wang, and Hung-yi Lee. 2024b. Developing instruction-following speech language model without speech instruction-tuning data. *arXiv preprint arXiv:2409.20007*.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Proc. NeurIPS*.

Ziyang Ma, Yakun Song, Chenpeng Du, Jian Cong, Zhuo Chen, Yuping Wang, Yuxuan Wang, and Xie Chen. 2025a. Language model can listen while speaking. *Proc. AAAI*.

Ziyang Ma, Guanrou Yang, Yifan Yang, Zhifu Gao, Jiaming Wang, Zhihao Du, Fan Yu, Qian Chen, Siqi Zheng, Shiliang Zhang, et al. 2024. An embarrassingly simple approach for LLM with strong ASR capacity. *arXiv preprint arXiv:2402.08846*.

Ziyang Ma, Guanrou Yang, Yifan Yang, Zhifu Gao, Jiaming Wang, Zhihao Du, Fan Yu, Qian Chen, Siqi Zheng, Shiliang Zhang, et al. 2025b. Speech recognition meets large language model: Benchmarking, models, and exploration. In *Proc. AAAI*.

Lingwei Meng, Shujie Hu, Jiawen Kang, Zhaoqing Li, Yuejiao Wang, Wenxuan Wu, Xixin Wu, Xunying Liu, and Helen Meng. 2024. Large language model can transcribe speech in multi-talker scenarios with versatile instructions. *arXiv preprint arXiv:2409.08596*.

Xingwei Qu, Yuelin Bai, Yinghao Ma, Ziya Zhou, Ka Man Lo, Jiaheng Liu, Ruibin Yuan, Lejun Min, Xueling Liu, Tianyu Zhang, et al. 2024. Mupt: A generative symbolic music pretrained transformer. *arXiv preprint arXiv:2404.06393*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. *Proc. ICML*.

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

Mohan Shi, Zengrui Jin, Yaoxun Xu, Yong Xu, Shi-Xiong Zhang, Kun Wei, Yiwen Shao, Chunlei Zhang, and Dong Yu. 2024. Advancing multi-talker asr performance with large language models. *Proc. SLT*.

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

Chen Wang, Minpeng Liao, Zhongqiang Huang, Jinliang Lu, Junhong Wu, Yuchen Liu, Chengqing Zong, and Jiajun Zhang. 2023a. Blsp: Bootstrapping language-speech pre-training via behavior alignment of continuation writing. *arXiv preprint arXiv:2309.00916*.

Jiaming Wang, Zhihao Du, Qian Chen, Yunfei Chu, Zhifu Gao, Zerui Li, Kai Hu, Xiaohuan Zhou, Jin Xu, Ziyang Ma, et al. 2023b. LauraGPT: Listen, attend, understand, and regenerate audio with GPT. *arXiv preprint arXiv:2310.04673*.

Jian Wu, Yashesh Gaur, Zhuo Chen, Long Zhou, Yimeng Zhu, Tianrui Wang, Jinyu Li, Shujie Liu, Bo Ren, Linquan Liu, et al. 2023a. On decoder-only architecture for speech-to-text and large language model integration. *Proc. ASRU*.

Shih-Lun Wu, Xuankai Chang, Gordon Wichern, Jeeweon Jung, François Germain, Jonathan Le Roux, and Shinji Watanabe. 2023b. BEATs-based audio captioning model with INSTRUCTOR embedding supervision and ChatGPT mix-up. *Proc. DCASE Challenge*.

Hongshen Xu, Zichen Zhu, Situo Zhang, Da Ma, Shuai Fan, Lu Chen, and Kai Yu. 2024a. Rejection improves reliability: Training llms to refuse unknown questions using rl from knowledge feedback. *Proc. COLM*.

Yaoxun Xu, Hangting Chen, Jianwei Yu, Qiaochu Huang, Zhiyong Wu, Shi-Xiong Zhang, Guangzhi Li, Yi Luo, and Rongzhi Gu. 2024b. SECap: Speech emotion captioning with large language model. *Proc. AAAI*.

Zhenhua Xu, Yujia Zhang, Enze Xie, Zhen Zhao, Yong Guo, Kwan-Yee K Wong, Zhenguo Li, and Hengshuang Zhao. 2024c. Drivegpt4: Interpretable end-to-end autonomous driving via large language model. *Proc. RA-L*.

Guanrou Yang, Ziyang Ma, Zhifu Gao, Shiliang Zhang, and Xie Chen. 2024a. Ctc-assisted llm-based contextual asr. *Proc. SLT*.

Guanrou Yang, Ziyang Ma, Fan Yu, Zhifu Gao, Shiliang Zhang, and Xie Chen. 2024b. Mala-asr: Multimedia-assisted llm-based asr. *Proc. Interspeech*.

Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, Chang Zhou, et al. 2024c. Air-bench: Benchmarking large audio-language models via generative comprehension. *Proc. ACL*.

Yuqing Yang, Ethan Chern, Xipeng Qiu, Graham Neubig, and Pengfei Liu. 2024d. Alignment for honesty. *Proc. NeurIPS*.

Gal Yona, Roee Aharoni, and Mor Geva. 2024. Can large language models faithfully express their intrinsic uncertainty in words? *Proc. EMNLP*.

Wenyi Yu, Changli Tang, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. 2024. Connecting speech encoder and large language model for ASR. *Proc. ICASSP*.

Ruibin Yuan, Hanfeng Lin, Yi Wang, Zeyue Tian, Shangda Wu, Tianhao Shen, Ge Zhang, Yuhang Wu, Cong Liu, Ziya Zhou, et al. 2024. Chatmusician: Understanding and generating music intrinsically with llm. *Proc. ACL*.

Hanning Zhang, Shizhe Diao, Yong Lin, Yi R Fung, Qing Lian, Xingyao Wang, Yangyi Chen, Heng Ji, and Tong Zhang. 2024. R-tuning: Teaching large language models to refuse unknown questions. *Proc. NAACL*.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2023. Multimodal chain-of-thought reasoning in language models. *Proc. TMLR*.

A Dataset Details

MMAU (Sakshi et al., 2024) is a novel benchmark designed to evaluate the capabilities of large-scale multimodal audio understanding models. MMAU consists of $10,000$ carefully curated audio-question-answer pairs, covering three major audio domains: speech, sound, and music. These questions involve both information extraction and reasoning tasks, spanning 27 distinct skills that challenge models to demonstrate advanced audio perception and domain-specific reasoning abilities. The dataset is divided into two parts: the Test-mini set, containing $1,000$ questions, and the main Test set, which includes $9,000$ questions. As the Test set is not open-sourced, we used the Test-mini set for our experiments. The Test-mini set reflects the same task distribution as the main Test set, and thus serves as a reliable evaluation set for reliable methods.

B Model Details

B.1 Introduction for different LALMs

SALMONN[3] (Tang et al., 2024) is one of the first universal LALMs capable of understanding and reasoning about speech, music, and general sounds. It employs a dual-encoder architecture, with the Whisper-Large-V2 (Radford et al., 2023) as the speech encoder and the BEATs (Chen et al., 2023) as the audio encoder. The outputs from both encoders are concatenated and processed by a window-level Q-Former (Li et al., 2023a) to align with the LLM Vicuna-13B (Chiang et al., 2023). The entire model was trained in three stages: the pre-training stage aimed at bridging the gap between audio encoders and the LLM, followed by instruction-tuning and activation-tuning stages to enhance the model's ability to follow human instructions and activate zero-shot emergent capabilities.

Qwen-Audio-Chat[4] (Chu et al., 2023) is a powerful LALM specifically designed to achieve universal audio understanding and facilitate flexible interaction based on human instructions. Based on Whisper-Large-V2 (Radford et al., 2023) and Qwen-7B (Bai et al., 2023), the model underwent a two-stage training process. In the first stage, a multi-task learning framework incorporating over 30 audio-related tasks was employed to endow the model with a comprehensive understanding of audio data. In the second stage, instruction-based fine-tuning was applied to enhance the model's ability to align with human intent, resulting in a strong interactive chat model.

Qwen2-Audio-Instruct[5] (Chu et al., 2024) represents the latest advancement in the Qwen-Audio series, capable of processing diverse audio inputs and providing either audio analysis or direct textual responses based on speech instructions. The model employs Whisper-Large-V3 (Radford et al., 2023) as its audio encoder and has undergone both supervised fine-tuning (SFT) and direct performance optimization (DPO) after pre-training, which has significantly enhanced its ability to follow complex instructions. Demonstrating strong performance across multiple benchmarks (Yang et al., 2024c; Li et al., 2024; Sakshi et al., 2024), Qwen2-Audio-Instruct is one of the most powerful open-source LALMs currently available.

B.2 Performance of different LALMs

We evaluated the performance of these powerful LALMs on the Test-mini set of MMAU. The baseline prompt in Table 4 and the IDK prompt in Table 5 are used to examine the effectiveness of the training-free method. As shown in Table 3, Qwen-Audio-Chat exhibits weak instruction-following capabilities, and adding the IDK prompt had little to no impact on accuracy, truthfulness, or reliability, potentially because the data used in the second stage was much smaller than the data used in the pre-training stage. In contrast, SALMONN demonstrated strong instruction-following abilities but was overly conservative. After adding the IDK prompt, the model's accuracy across all three audio modalities significantly decreased, while its truthfulness notably increased, indicating a strong inclination to refuse to answer questions. We hypothesize that this over-strong instruction-following ability is related to the activation tuning in the third stage of SALMONN's training. Qwen2-Audio-Instruct outperforms other models on most of the evaluation metrics, for which it is chosen as the baseline for our main experiments.

[3] https://huggingface.co/tsinghua-ee/SALMONN/blob/main/salmonn_v1.pth

[4] https://huggingface.co/Qwen/Qwen-Audio-Chat

[5] https://huggingface.co/Qwen/Qwen2-Audio-7B-Instruct

Table 3: *Accuracy* (Acc%↑), *Truthfulness* (Tru%↑), and *Reliability* (Rel%↑) performance comparison of Qwen-Audio-Chat, SALMONN, and Qwen2-Audio-Instruct on the MMAU benchmark across sound, speech, and music modalities. Both the baseline and the IDK Prompting approaches were evaluated, with GPT answer normalization applied. The best-performing items are highlighted in **bold**, and the second-best items are underlined.

Models	IDK Prompting	Sound			Music			Speech			Total		
		Acc%	Tru%	Rel%	Acc%	Tru%	Rel%	Acc%	Tru%	Rel%	Acc%	Tru%	Rel%
Qwen-Audio-Chat	✗	57.66	58.86	58.84	53.29	53.59	53.59	35.44	35.74	35.73	48.80	49.40	49.40
Qwen-Audio-Chat	✓	53.75	55.26	55.23	53.29	53.59	53.59	37.24	37.54	37.54	48.10	48.80	48.80
SALMONN	✗	50.46	52.25	52.22	49.70	50.00	50.00	30.63	37.54	37.06	43.60	46.60	46.51
SALMONN	✓	28.53	**87.69**	52.69	17.96	**85.03**	40.05	8.41	**92.79**	21.59	18.30	**88.50**	39.22
Qwen2-Audio-Instruct	✗	**60.96**	60.96	60.96	**55.09**	55.09	55.09	**50.75**	50.75	50.75	**55.60**	55.60	55.60
Qwen2-Audio-Instruct	✓	58.26	76.28	**73.03**	54.19	66.77	**65.19**	43.84	58.26	**56.18**	52.10	67.10	**64.85**

C Prompting Details

C.1 Prompt Template for LALM

Given {*Audio*} and {*Question*}, we use some templates to generate unnormalized answers (which means that they cannot be used directly for evaluation processing). For the Baseline and LoRA Fine-tuning (Table 4), IDK Prompting (Table 5), and MCoT Prompting (Table 6), LALM only needs to be inferred once, while for Task Agent, the model needs to be inferred multiple times. The specific templates are shown in the tables bellow.

Baseline

Input:
{*Audio*} {*Question*} Select one option from the provided choices:
{*Content_of_A*}
{*Content_of_B*}
{*Content_of_C*}
{*Content_of_D*}
Output:
{*Answer*}

Table 4: The Prompt Template for the baseline on MMAU.

C.2 Prompt Template for Answer Normalization

Although LALM is required to output a unique option, the output is likely diverse due to limited instruction-following capability of LALM. Therefore, we use the *gpt-4o-mini* API to further normalize the answer. The corresponding prompt is shown in Table 8.

IDK Prompting

Input:
{*Audio*} {*Question*} Select one option from the provided choices:
{*Content_of_A*}
{*Content_of_B*}
{*Content_of_C*}
{*Content_of_D*}
Output 'IDK' if you don't know the answer.
Output:
{*Answer*}

Table 5: The Prompt Template for IDK Prompting on MMAU.

MCoT Prompting

Input:
{*Few-shot Examples*}
{*Audio*} {*Question*} Select one option from the provided choices:
{*Content_of_A*}
{*Content_of_B*}
{*Content_of_C*}
{*Content_of_D*}
Let's think step by step.
You can first analyze the sound, music, or speech and then answer the question.
Output 'IDK' if you don't know the answer.
Output:
{*Answer*}

Table 6: The Prompt Template for MCoT Prompting on MMAU.

Task Agent

Input:
{*Audio*} Identify the type of audio. Select one option from the provided choices:
Sound
Music
Speech
Output:
{*Type*}

Input:
{*Audio*} What is the {*Type*} content?
Output:
{*Content*}

Input:
{*Audio*} The {*Type*} content is: {*Content*}
{*Question*} Select one option from the provided choices:
{*Content_of_A*}
{*Content_of_B*}
{*Content_of_C*}
{*Content_of_D*}
Output 'IDK' if you don't know the answer.
Output:
{*Answer*}

Table 7: The Prompt Template for Task Agent on MMAU.

Answer Normalization

Input:
According to the answer, select one option from the provided choices.
The answer is: {*Answer*}
The choices are:
{*Content_of_A*}
{*Content_of_B*}
{*Content_of_C*}
{*Content_of_D*}
IDK
Don't output any other information.
Output:
{*Answer (Normalized)*}

Table 8: The Prompt Template for answer normalization on With OpenAI API.

D Training Details

Table 9 shows the hyper-parameters of SFT in different audio modalities, including learning rate, LoRA alpha, LoRA rank, and LoRA target modules.

Table 9: Hyper-parameters with LoRA Fine-tuning for each modality on the MMAU dataset.

	Sound	Music	Speech
learning rate	3×10^{-5}		
LoRA alpha	32	16	
LoRA rank	8		
target modules	{k_proj, q_proj, v_proj}		

E Metric Details

Here we analyze the condition for an invalid Reliability metric. In the case of a vanilla LALM, suppose that the model simply answers questions based on its existing knowledge with an accuracy of α, where $0 \leq \alpha \leq 1$. As Equation 5,the original reliability Rel_{org} of the model can be computed as:

$$\begin{aligned} Rel_{org} &= 0 \cdot \alpha + 1 \cdot \alpha \\ &= \alpha \end{aligned} \tag{11}$$

Now, consider the case where a reliable method is applied. Let Δ_{Con} and Δ_{Hum} represent the increase in conservativeness and humbleness, respectively, as per Equations 8 and 9. The lower bound of a valid reliable method occurs when Δ_{Con} and Δ_{Hum} are at the same ratio ρ. The resulting Accuracy, Rejection Rate, Truthfulness, and Reliability can be computed as follows:

$$\begin{aligned} Acc_{new} &= (1 - \rho)\alpha \\ &= \alpha - \rho\alpha \end{aligned} \tag{12}$$

$$\begin{aligned} Rej_{new} &= \rho\alpha + \rho(1 - \alpha) \\ &= \rho \end{aligned} \tag{13}$$

$$\begin{aligned} Tru_{new} &= Acc + Rej \\ &= \alpha + \rho - \rho\alpha \end{aligned} \tag{14}$$

$$\begin{aligned} Rel_{new} &= Rej \cdot Acc + (1 - Rej) \cdot Tru \\ &= \rho(\alpha - \rho\alpha) + (1 - \rho)(\alpha + \rho - \rho\alpha) \\ &= \alpha - \rho\alpha + \rho - \rho^2 \end{aligned} \tag{15}$$

Table 10: *Accuracy* (Acc%↑), *Truthfulness* (Tru%↑), and *Reliability* (Rel%↑) performance of Qwen2-Audio-Instruct with LoRA Fine-tuning on different modalities.

Training Modality	Sound			Music			Speech		
	Acc%	Tru%	Rel%	Acc%	Tru%	Rel%	Acc%	Tru%	Rel%
Sound	-	-	-	46.71	73.05	66.11	48.95	63.66	61.50
Music	62.76	70.57	69.96	-	-	-	46.85	60.06	58.31
Speech	60.66	72.97	71.46	55.99	68.26	66.76	-	-	-

Table 11: *Relative Conservativeness Increase* ($\Delta_{Con}\%$ ↓), *Relative Humbleness Increase* ($\Delta_{Hum}\%$ ↑) and *Reliability Gain Index* (RGI↑) performance of Qwen2-Audio-Instruct with LoRA Fine-tuning on different modalities.

Training Modality	Sound			Music			Speech		
	$\Delta_{Con}\%$	$\Delta_{Hum}\%$	RGI	$\Delta_{Con}\%$	$\Delta_{Hum}\%$	RGI	$\Delta_{Con}\%$	$\Delta_{Hum}\%$	RGI
Sound	-	-	-	15.57	23.95	0.19	12.31	21.02	0.23
Music	6.31	14.41	0.36	-	-	-	10.81	15.32	0.15
Speech	7.51	16.82	0.35	9.88	18.26	0.27	-	-	-

We now analyze when the reliability after applying the method exceeds the original reliability, which leads to the following inequality:

$$\begin{aligned} &Rel_{new} > Rel_{org} \\ \Rightarrow &\alpha - \rho\alpha + \rho - \rho^2 > \alpha \\ \Rightarrow &\rho < 1 - \alpha \end{aligned} \tag{16}$$

where the Reliability metric does not accurately describe the nature of what it expresses, because ineffective reliable methods increase the Reliability metric if the value of ρ less than $1 - \alpha$ is satisfied.

For an illustrative example, consider a model initially producing 50% correct and 50% incorrect answers. By applying Equation 5, we can calculate the Reliability of the original unreliable model as:

$$Rel = 0 \times 50\% + 1 \times 50\% = 50\%. \tag{17}$$

After applying a reliable method, if 10% of both the correct and incorrect answers are converted into rejections, the new Accuracy/Rejection/Error rate would be 40%/20%/40%, respectively. In this case, the reliable method would appear ineffective because, after applying the method, the distribution of correct and incorrect answers turning into rejections is similar to what would occur with random sampling. However, the reliability increases to:

$$Rel = 20\% \times 40\% + 80\% \times 60\% = 56\%, \tag{18}$$

indicating an ineffective measurement of the model's reliability. Here, the increase in reliability is deceptive, as it results from indiscriminate rejection rather than true improvement. The RGI, introduced in the main text, addresses this issue by comparing the relative increase in humbleness and conservativeness.

F More results

F.1 LoRA Fine-tuning Results on Different Modalities

Table 10 shows the LoRA fine-tuning performance on the Accuracy, Truthfulness, and Reliability of Qwen2-Audio-Instruct trained on one modality and tested on other modalities, while Table 11 shows the LoRA fine-tuning performance on the Relative Conservativeness Increase, Relative Humbleness Increase, and RGI.

F.2 Rejection Rate on Different Modalities

Table 12 shows the proportion of IDK items in the IDK training dataset and the Rejection Rate tested on different audio modalities. The hyperparameters come from Table 9.

Table 12: Rejection Rate on different modalities

Training Modality	Rejection Rate (%)			
	IDK Dataset	Sound	Music	Speech
Sound	63.06	37.24	26.35	14.71
Music	61.98	7.81	34.73	13.21
Speech	65.47	12.31	12.28	56.76