Beyond Accuracy: ARIA-Rubrics for Evaluating Audio Reasoning in Large Audio Language Models

Yupei Li
Imperial College London, UK
yl7622@ic.ac.uk

Qiyang Sun
Imperial College London, UK
q.sun23@imperial.ac.uk

Mohamed Mady
Technische Universität München
München, Germany
mohamed.mady@tum.de

Chenxi Wang
Mohamed bin Zayed University of Artificial Intelligence
Abu Dhabi, AE
chenxi.wang@mbzuai.ac.ae

Zhengwei Gong
Shanghai Jiao Tong University, Shanghai, CN
gzwsjtu98@sjtu.edu.cn

Berrak Sisman
Johns Hopkins University, Maryland, US
sisman@jhu.edu

Björn Schuller
Technische Universität München, München, Germnay
schuller@tum.de

Abstract

Large Audio Language Models (LALMs) have shown strong performance on audio reasoning benchmarks, but accuracy alone cannot distinguish true reasoning from superficial pattern matching, often overestimating reasoning ability since high scores may result from guessing rather than genuine audio understanding. Evaluating the reasoning process itself is essential for improving LALMs' reasoning ability, yet remains challenging. Existing methods either rely on costly human annotation or opaque LLM-as-judge approaches, making them impractical, biased, and lacking transparency. Moreover, audio reasoning introduces unique challenges absent in text-based settings, perceptual hallucination and cross-modal alignment between audio understanding and textual inference, hence text-based evaluation frameworks cannot be directly applied. Therefore, we propose ARIA-Rubrics (Audio Reasoning Integrity Assessment), a lightweight, annotation-free gold reasoning chains, automatic and transparent framework comprising six complementary metrics that evaluate audio reasoning quality across perceptual grounding, reasoning coherence, and answer consistency. We use Chain-of-Thought prompting as an externalization mechanism to make the reasoning process observable. Experiments on 9 models across 2 benchmarks identify three reasoning modes of current LALMs with actionable directions for future development, with ARIA-Rubrics achieving high correlation with human judgments. The code is available at the GitHub Repository

1 Introduction

Reasoning is a fundamental yet challenging capability for large language models (LLMs) (Plaat et al., 2025), and becomes even more demanding for large audio language models (LALMs), where additional acoustic modalities need to be perceived, aligned, and integrated into the thinking process (Wang et al., 2024). Several benchmark datasets have been developed to assess LALMs' reasoning ability, such as MMAR (Ma et al., 2026b) and MMAU (Sakshi et al., 2025). To tackle these challenging benchmarks, a range of models have been developed using diverse pre-training backbones combined with post-training strategies including supervised fine-tuning (Das et al., 2024), reinforcement learning (Wu et al., 2026), and chain-of-thought (CoT) training (Ma et al., 2025). While existing models reviewed in Section 2 have shown great potential (Su et al., 2025), current evaluations rely primarily on accuracy or task completion rates, which provide a biased view of reasoning capabilities and cannot reveal whether models genuinely reason or merely exploit statistical patterns, a distinction crucial for understanding their limitations.

Therefore, assessing the reasoning process itself is essential, as it can uncover critical failure modes (Shojaee et al., 2026) and guide future model improvements: *Reasoning*, where the CoT faithfully

Findings of the Association for Computational Linguistics: IJCNLP-AACL 2026, pages 1560–1581
November 6-10, 2026 ©2026 Association for Computational Linguistics

drives the derivation of the correct answer; *Scaffolding*, where the reasoning appears structurally complete but is vacuous or irrational, failing to meaningfully support the correct answer; and *Decoration*, where the reasoning is entirely inconsistent and has no causal effect on the final answer. Existing approaches to process-level evaluation reviewed in Section 2 fall into two categories: human annotation (Yang et al., 2025; Huang et al., 2025), which is costly and labour-intensive at scale, and LLM-as-judge methods (Bi et al., 2026), which suffer from opacity and limited reliability.

Textual LLMs have inspired evaluation methods beyond human annotation and LLM judges. For example, Golovneva et al. (2023) proposed multi-dimensional metrics for semantic consistency and coherence, while Prasad et al. (2023) measured the informativeness of each reasoning step. However, these approaches do not directly apply to audio reasoning, which requires perceptual grounding: the quality of acoustic perception underlies all subsequent reasoning, and fabricated or inaccurate observations cannot be detected by text-only metrics. These limitations motivate the design of audio-native evaluation dimensions.

Building on the aforementioned limitations, we propose ARIA-Rubrics (Audio Reasoning Integrity Assessment) to systematically examine the intermediate reasoning process. Using CoT to reveal how models reach their answers, we design six complementary metrics covering acoustic perception, reasoning step progressiveness, and conclusion consistency. A combination of automatic metrics, lightweight model scoring, and targeted LLM judgments balances transparency and effectiveness, reducing dependence on opaque black-box judgments while leveraging natural language understanding where needed. Experiments on 9 models across MMAR and MMAU-mini reveal the three reasoning modes described above, providing diagnostic insights and actionable guidance for future development, with reliability confirmed through human evaluation.

In summary, our **contributions** are as follows. First, we propose **annotation-free gold reasoning chains, transparent, and automatic ARIA-Rubrics**, the first audio reasoning quality evaluation framework to the best of our knowledge, comprising six carefully designed complementary metrics covering both audio-native perceptual grounding and reasoning process quality. Second, we conduct comprehensive experiments on 9 models across 2 benchmarks, identifying three reasoning modes of current LALMs, *Reasoning*, *Scaffolding*, and *Decoration*, and providing actionable directions for future model development. Third, we validate ARIA-Rubrics through human evaluation, demonstrating strong correlation with human judgments and confirming the reliability of our automated metrics.

2 Related work

LALMs have rapidly advanced with diverse architectures and training paradigms in audio reasoning. Pretrained models such as Qwen2-Audio (Yang et al., 2024a), Qwen2.5-Omni (Xu et al., 2025), Phi-4 Multimodal (Abouelenin et al., 2025), and Gemma-3n (Team et al., 2025) provide strong baselines through large-scale audio-text pretraining, while post-training strategies including reinforcement learning (R1-AQA (Wu et al., 2026)) and structured CoT fine-tuning (Audio-Reasoner (Ma et al., 2025), AudSemThinker (Wijngaard et al., 2026)), further boost performance. Closed-source models such as GPT-4o Audio[1] and Gemini-2.5 Flash(Comanici et al., 2025) represent the commercial frontier. These models are evaluated on benchmarks including MMAU (Sakshi et al., 2025), MMAR (Ma et al., 2026b), AudioBench (Wang et al., 2025b), MUSE (Carone et al., 2026), and AIR-Bench (Yang et al., 2024b), covering diverse audio understanding and reasoning tasks. However, since these benchmarks rely solely on accuracy or task completion, they cannot distinguish genuine reasoning from statistical pattern exploitation. They reflect answer correctness, but not the quality or faithfulness of the reasoning process.

Existing frameworks for audio reasoning evaluation include MMAR-Rubrics (Ma et al., 2026a), which relies on proprietary models and human-annotated gold CoTs, limiting reproducibility and scalability; CAFE (Mao et al., 2026), which uses LLM-as-judge to quantify perception and utilization errors but remains a black-box approach; and OCRA (Sedláček et al., 2025), which depends entirely on human annotators without detailed guidance. In contrast, ARIA-Rubrics offers a lightweight, transparent, annotation-free gold reasoning chains framework with audio-native metrics, requiring neither large LLMs nor human annotations.

[1] https://developers.openai.com/api/docs/models/gpt-4o-audio-preview

Evaluating reasoning chain quality has been widely studied in NLP. CoT prompting (Wei et al., 2022) externalizes intermediate reasoning steps, making them observable and assessable, and their quality has been shown to correlate strongly with final answer correctness (Lightman et al., 2024), establishing CoT as a reliable proxy for reasoning ability. Accordingly, many evaluation rubrics leverage CoT. For instance, ROSCOE (Golovneva et al., 2023) measures semantic consistency and coherence between steps, ReCEval (Prasad et al., 2023) assesses correctness and informativeness via natural language inference, Process Reward Models (Zheng et al., 2025) verify step-level mathematical reasoning using human-annotated supervision, FaithScore (Jing et al., 2024) evaluates faithfulness to source context through atomic fact verification, and SocREval (He et al., 2024) applies the Socratic method for annotation-free gold reasoning chains reasoning assessment. These automatic metrics avoid costly human annotation. However, these frameworks are designed for text-only settings and lack perceptual grounding dimensions, and perceptual hallucinations cannot be detected by text-based metrics, but the aspects motivate the design of audio-native evaluation dimensions in ARIA-Rubrics.

3 ARIA-Rubrics framework

3.1 Overview

ARIA-Rubrics is designed to evaluate the reasoning quality of LALMs by examining the intermediate reasoning process rather than the final answer alone. The core insight is that a faithful reasoning process should be grounded in accurate acoustic perception, progress coherently toward the answer, and maintain consistency between reasoning and conclusion. To make the reasoning process observable, we adopt CoT prompting as an externalization mechanism, eliciting models to produce structured reasoning chains rather than direct answers. Given this structured output, ARIA-Rubrics evaluates six complementary metrics (M1–M6) targeting different dimensions of reasoning quality. The overall pipeline is illustrated in Figure 1. The final ARIA score is computed as the mean of all six metrics.

3.2 CoT prompt design

As previously stated, we design a structured four-step CoT prompt comprising *Perception*, *Analysis*, *Reasoning*, and *Answer*. Each step serves a distinct cognitive function: *Perception* requires the model to enumerate the sound entities it actually hears and describe their acoustic properties and temporal order without interpretation; *Analysis* connects acoustic observations to their possible meanings without drawing final conclusions; *Reasoning* weighs the evidence and eliminates alternatives to arrive at a conclusion; and *Answer* states the final choice, with the actual prompt shown in Appendix A.1.

We design the four-step structure by examining LLM-generated CoT annotations from ALR (Diao et al., 2025) and CoTA (Xie et al., 2025), identifying a consistent cognitive flow from acoustic perception to semantic inference as their core reasoning pattern, and human-annotated CoT examples remain unavailable in open-source form (e.g., MMAR (Ma et al., 2026b)). This design choice is further supported by Ma et al. (2025), who showed that different CoT template variants, including zero-shot, manual few-shot, and other prompt formats, yield accuracy differences within 3% absolute, suggesting that prompt engineering alone does not significantly constrain inherent reasoning capabilities in a biased assessment framework. We further validate our specific prompt design through a sanity check comparing model accuracy with and without our CoT prompt across all evaluated models, reported in Appendix B.1. Results show no huge performance difference in most cases, confirming that our structured prompt does not restrict model reasoning ability, and that observed accuracy degradation under CoT prompting more likely reflects the model's lack of genuine reasoning ability rather than prompt interference.

3.3 Six Complementary Metrics

3.3.1 Acoustic Grounding Score (M1)

Since perception is a fundamental component of audio reasoning, we introduce the first metric specifically designed to evaluate acoustic perception ability. The Acoustic Grounding Score (AGS) measures whether the sound entities identified during the *Perception* stage are genuinely grounded in the input audio signal. This metric detects perceptual hallucinations, where models fabricate nonexistent acoustic events or misidentify entities. For each entity e_i listed in the *Entities* line of *Perception*, we compute its similarity to the whole audio signal using a Contrastive Language-Audio Pretraining (CLAP) model (Wu* et al., 2023), a pre-trained model to align audio and text representations in a

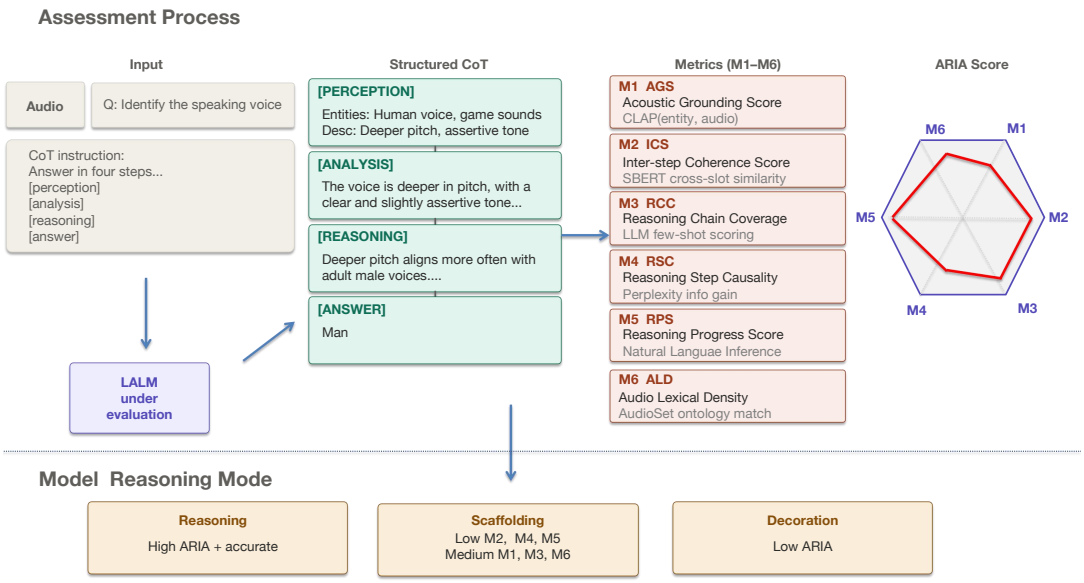


Figure 1: Overview of the ARIA-Rubrics evaluation pipeline. Given an audio input and a question, the candidate LALM is prompted with a structured four-step CoT template to externalize its reasoning process. The resulting CoT output is evaluated by six complementary metrics (M1–M6) covering hallucination detection, cross-slot coherence, content substantiveness, reasoning progress, reasoning–answer consistency, and audio vocabulary density. A final ARIA score is computed as the mean of all six metrics. Based on the score profile, each model could be categorised into one of three reasoning modes: *Reasoning*, *Scaffolding*, or *Decoration*.

shared embedding space. The AGS is defined as:

$$\text{AGS} = \frac{1}{|E|} \sum_{i=1}^{|E|} \text{CLAP_sim}(e_i, \mathbf{a}), \qquad (1)$$

where E is the set of extracted entities and $\mathbf{a}$ is the whole audio input. It also prevents any hack strategies in which models output as many sound entities as possible to artificially increase the average score. If no entities are detected in *Perception*, M1 is 0.

3.3.2 Inter-step Coherence Score (M2)

After the perception stage is recognized, a reliable reasoning model should correctly utilize the extracted evidence for subsequent reasoning. The Inter-step Coherence Score (ICS) evaluates semantic coherence between consecutive reasoning steps, assessing whether each step is grounded by prior context. Specifically, coherence is computed across the first three reasoning steps (excluding the final answer, which is a single output) using sentence-level embeddings (Song et al., 2020) (*all-mpnet-base-v2*), chosen for its strong performance on downstream text understanding and support for long-context embeddings. The ICS is defined as:

$$\text{ICS} = \tfrac{1}{2}\big(c(f(A), f(P)) + c(f(R), f(A))\big) \qquad (2)$$

where $f(\cdot)$ denotes the sentence embedding function, $c(\cdot)$ represents cosine similarity, P, A, R correspond to the sentences from the *Perception*, *Analysis*, and *Reasoning* sections, respectively. If either slot is absent, M2 is 0.

3.3.3 Reasoning Chain Coverage (M3)

Though the references from previous steps can be verified, we also need to assess whether each step in the CoT contains substantive content rather than meaningless contents. Accordingly, the Reasoning Chain Coverage (RCC) is computed as follows: for each slot $s \in \{\textit{Perception}, \textit{Analysis}, \textit{Reasoning}\}$, we employ Qwen2.5-1.5B-Instruct as a lightweight scorer with few-shot prompting to assign a substantiveness score $\sigma(s) \in [0, 1]$, where 0 corresponds to entirely empty or repetitive content, and 1 corresponds to highly specific and informative content. The prompt for the scorer is shown in Appendix B.2. The *Answer* slot is scored deterministically: 1.0 if it is present and non-empty, and 0.0 otherwise. The RCC is computed as the average of four slot scores. If a slot is absent, its corresponding score ($\sigma(P/A/R/Ans)$, where Ans denotes sentences from the *Answer* section) is set to 0.

We adopt an LLM-as-a-judge strategy, noting that the task is relatively simple and mainly requires basic language understanding. Few-shot examples in the evaluation prompts improve re-

liability, proved by Zhang et al. (2023). Unlike fully black-box LLM-based evaluations of overall reasoning ability, our framework leverages LLMs' strengths in language understanding while avoiding tasks where their judgments are unreliable.

3.3.4 Reasoning Step Causality (M4)

The reasoning process should be complete and coherent, and also progressively guide toward the correct answer. Reasoning Step Causality (RSC) measures whether each step advances toward the ground-truth answer by estimating the information gain contributed by that step. Specifically, we track how the probability of the correct answer changes as each step is added to the context, using a small language model to compute conditional perplexity. If an added step makes the correct answer more predictable, the perplexity decreases; we define information gain as the reduction in perplexity relative to the previous context. To ensure a non-negative score, we apply a sigmoid transformation to the information gain, yielding the exponential form below. This design follows ROSCOE (Golovneva et al., 2023), which assesses step-level informativeness without human-annotated references. We use Qwen2.5-1.5B-Instruct (as in M3) as the evaluator to avoid reliance on additional pretrained models.

$$\begin{aligned}\text{Info}_A &= \text{PPL}(\hat{Ans} \mid P, A) - \text{PPL}(\hat{Ans} \mid P)\\ \text{Info}_R &= \text{PPL}(\hat{Ans} \mid P, A, R)\\ &\quad - \text{PPL}(\hat{Ans} \mid P, A)\\ \text{RSC} &= \frac{1}{2}\sum_{i\in\{A,R\}} \frac{1}{1+e^{\text{Info}_i}}\end{aligned} \tag{3}$$

where $\hat{Ans}$ is the groundtruth Answer. If a slot is absent, the corresponding score ($\text{Info}_{A/R}$) is 0.

3.3.5 Reasoning Progress Score (M5)

Collecting correct information is insufficient if the model does not progress fluently. The Reasoning Progress Score (RPS) evaluates whether each step is conditioned on the previous one, distinguishing it from M2, which measures whether prior context is referenced. We adopt Natural Language Inference (NLI), a well-studied task for determining whether a premise entails a hypothesis, as a natural fit for this metric. We use Roberta-large-mnli (Liu et al., 2019) fine-tuned on NLI tasks as the classifier. The RPS is calculated as:

$$\text{RPS} = \frac{1}{3}\big(g(P,A) + g(A,R) + g(R,Ans)\big), \tag{4}$$

where $g(\cdot)$ is the NLI model's entailment score. If a slot is absent, its corresponding score is 0.

3.3.6 Audio Lexical Density (M6)

After evaluating all reasoning processes, a common criticism of LLMs is that their reasoning chains are often overly long (An et al., 2026), potentially due to mechanisms such as self-consistency checks (Wang et al., 2023) or exploration of multiple solution paths (Wang et al., 2025a). However, excessive length may cause models to lose focus or produce unnecessarily verbose reasoning. We propose Audio Lexical Density (ALD), which measures the proportion of audio-relevant vocabulary in the reasoning chain. We hypothesize that, in audio reasoning tasks, greater use of audio-related terms correlates with more meaningful inference. Although this metric is an approximation, it can effectively penalize excessive self-reflection without progress and the inclusion of irrelevant content.

Specifically, we utilize the AudioSet ontology (Gemmeke et al., 2017), which contains a comprehensive set of audio-related terms, including verbal actions, sound-producing subjects, and environmental sounds. We deem these are complete and enough as these are annotated by human experts. The audio reasoning benchmarks mainly concern these domains. We manually filtered out some synonymous terms (e.g., baby and infant), leaving a total of 417 words (listed in Appendix C). For each word in the CoT output, we compute its similarity to each ontology term using the word2vec model[2] (Mikolov et al., 2013), selecting the maximum similarity as the score for that word. The overall ALD score is then obtained by averaging these maximum similarities across all words in the reasoning chain.

3.4 Overall discussion

The six metrics are designed to be complementary rather than redundant, each targeting a distinct dimension of reasoning quality while collectively preventing gaming through mutual constraints. Specifically, artificially inflating any single metric will inevitably suppress others: a model that fabricates plausible-sounding entities to boost M6 will be penalised by M3 if those entities are repetitively listed; a model that rushes to the correct answer too quickly to maximise M5 will exhibit low M1 and M4 due to insufficient perceptual grounding; and a model recalling too much to increase M2 will lead

[2] https://huggingface.co/fse/word2vec-google-news-300

to decrease in M5. Together, the six metrics cover the full reasoning pipeline from acoustic perception to final answer, and the ARIA score is computed as their unweighted mean, yielding a value $\in [0, 1]$.

The metric profile also provides a diagnostic basis for identifying the three reasoning modes. Models exhibiting high accuracy but low ARIA scores are characterised as *Decoration*. Conversely, high ARIA with high accuracy indicates genuine *Reasoning*. The *Scaffolding* mode is primarily identified through low scores on M2, M4, and M5, which assess the semantic content and meaningfulness of the reasoning process; a model may maintain structural coherence (high M3) while producing vacuous or disconnected content that fails these meaning-oriented checks.

4 Experiments and Results

We evaluate ARIA-Rubrics on two audio reasoning benchmarks: MMAU-mini (Sakshi et al., 2025) and MMAR (Ma et al., 2026b), each comprising 1,000 multiple-choice questions with no sample overlap between the two datasets. To validate the effects of our rubrics, we select 9 LALMs spanning diverse base models and training paradigms, covering both open-source and closed-source systems. Detailed hyperparameter settings and implementation details are provided the Appendix D. Results are presented in Table 1. We extract final answers from model outputs via a three-stage matching procedure: exact string matching against answer choices, substring matching, and letter-to-choice index mapping. The full details are provided in Appendix E.

We observe **a systematic divergence between ARIA scores and accuracy rankings**, indicating that accuracy alone provides an incomplete picture of reasoning quality. Two representative cases illustrate this (Table 3): First, a model may achieve high accuracy despite low ARIA scores; for example, AudSemThinker (M3=0.36, M5=0.50) answers correctly, but its CoT lacks acoustic grounding: the Perception step only assigns an accent label without phonetic evidence, and the Reasoning step circularly restates this label without inferential progression. Second, a model may score high on ARIA despite an incorrect answer; Qwen2.5-Omni (M3=0.84, M5=1) demonstrates genuine acoustic observation and logical reasoning, yet fails due to a knowledge gap about Indian English phonetics rather than reasoning failure. These cases show that ARIA provides diagnostic insights invisible to accuracy-based evaluation, enabling a faithful assessment of model reasoning ability, while focusing on reasoning rather than knowledge coverage.

We additionally conducted an LLM-as-a-judge evaluation (Gu et al., 2026) to examine whether there is any divergence between ARIA scores and LLM-based judgements (Deepseek-v4-pro (Xu et al., 2026) and Gemini-v3-flash (Team et al., 2023)), with prompts of LLM-as-a-judge in Appendix B.3. The results in Table 2 show that the ranking produced by ARIA is largely consistent with the overall model performance assessed by the LLM judge, while providing more interpretable insights into specific directions for future improvement. Moreover, unlike LLM-based evaluation, which can be computationally and token intensive, ARIA provides a more efficient and structured evaluation framework.

Model reasoning modes We further analyze model reasoning modes, providing additional validation of our rubrics. Specifically, we manually perturb CoTs on randomly selected instances, injecting two types of altered reasoning: a Decoration CoT, with unstructured and vacuous content (e.g., random words like 0*I love dog*), and a Scaffolding CoT, with correct structure but fabricated content (e.g., *the sound is alike bird and tiger*). We then combine the perturbed CoT with the question and audio to assess whether each model's final answer changes. Full prompts and responses are in Appendix B.4, with results summarized in Table 4.

The results show a strong alignment between perturbation-derived mode assignments and ARIA rankings: Reasoning models rank highest (1–3), Scaffolding models mid (5–7), and Decoration models lowest, confirming that ARIA reflects reasoning quality. Closer examination reveals distinct failure patterns: Scaffolding models, such as AudSemThinker, produce structurally complete but often irrational CoTs (e.g., "low-pitched hooting $\rightarrow$ tiger"), occasionally reaching correct answers, consistent with low M2, M4, and M5 scores. Decoration models (R1-AQA, Qwen2-Audio) show no causal effect of CoT on the final answer. In contrast, Reasoning models resist misleading CoTs and reground their answers in the audio, demonstrating genuine cross-modal reasoning.

Dataset-independent analysis for ARIA rubrics We apply ARIA rubrics across different reason-

Table 1: Results on MMAR and MMAU-mini benchmarks. Cell color indicates per-column rank (darker = higher rank, 9-level warm gradient). ARIA = mean of M1–M6. Acc means accuracy.

Model	M1		M2		M3		M4		M5		M6		ARIA	Acc(%)
MMAR														
Open-source Models														
Qwen2-Audio	0.1963	±0.1988	0.6169	±0.1917	0.6967	±0.1711	0.5220	±0.2757	0.2015	±0.1792	0.4326	±0.0815	0.4443	17.7%
Qwen2.5-Omni	0.4344	±0.1722	0.6815	±0.2202	0.7561	±0.1987	0.5887	±0.2734	0.2873	±0.2160	0.5215	±0.1238	0.5449	44.3%
Gemma-3n-E2B	0.2861	±0.1754	0.6442	±0.1972	0.7362	±0.1720	0.6315	±0.2934	0.3215	±0.2140	0.4356	±0.0944	0.5092	27.3%
R1-AQA	0.2107	±0.1900	0.5700	±0.2004	0.6891	±0.1589	0.5358	±0.2764	0.1812	±0.1699	0.4237	±0.0779	0.4351	21.6%
Audio-Reasoner	0.2401	±0.1747	0.6159	±0.2584	0.6293	±0.2369	0.6099	±0.2191	0.2887	±0.2344	0.4817	±0.1202	0.4776	20.1%
AudSemThinker	0.3309	±0.1923	0.4626	±0.3480	0.5735	±0.3408	0.5672	±0.2366	0.2317	±0.2313	0.6171	±0.2106	0.4638	31.1%
Phi-4	0.3432	±0.1665	0.5972	±0.2820	0.6481	±0.2538	0.5129	±0.2499	0.2210	±0.2029	0.6143	±0.1497	0.4895	27.1%
Closed-source Models														
GPT-4o-audio	0.3725	±0.1686	0.7142	±0.0849	0.7838	±0.0589	0.7506	±0.2522	0.4346	±0.2057	0.4219	±0.0608	0.5796	67.6%
Gemini 2.5 Flash	0.3808	±0.1630	0.6776	±0.0908	0.7729	±0.0730	0.7426	±0.2379	0.4021	±0.1938	0.4375	±0.0557	0.5689	62.6%
MMAU-mini														
Open-source Models														
Qwen2-Audio	0.1941	±0.1315	0.6123	±0.2075	0.6959	±0.1871	0.5346	±0.2180	0.2209	±0.1873	0.4462	±0.0559	0.4507	24.1%
Qwen2.5-Omni	0.1929	±0.1305	0.6654	±0.1958	0.7587	±0.1681	0.6364	±0.2624	0.3481	±0.1903	0.4771	±0.0705	0.5131	58.4%
Gemma-3n-E2B	0.3174	±0.1764	0.6207	±0.1908	0.7181	±0.1655	0.6903	±0.2784	0.3237	±0.2064	0.4491	±0.0943	0.5199	36.5%
R1-AQA	0.2670	±0.1815	0.5847	±0.2072	0.6926	±0.1737	0.5536	±0.2775	0.2056	±0.1653	0.4490	±0.0898	0.4588	26.3%
Audio-Reasoner	0.3032	±0.1636	0.5774	±0.2946	0.6186	±0.2646	0.6877	±0.2029	0.2860	±0.2334	0.5097	±0.1622	0.4971	34.0%
AudSemThinker	0.3637	±0.1764	0.5237	±0.3325	0.6178	±0.3120	0.5476	±0.2286	0.2616	±0.2239	0.6046	±0.1973	0.4865	46.0%
Phi-4	0.3443	±0.1724	0.6507	±0.2004	0.7300	±0.1650	0.6255	±0.2730	0.2744	±0.1918	0.5521	±0.1090	0.5295	40.7%
Closed-source Models														
GPT-4o-audio	0.4035	±0.1632	0.6836	±0.1194	0.7780	±0.0969	0.7491	±0.2392	0.4161	±0.1975	0.4502	±0.0752	0.5801	55.8%
Gemini 2.5 Flash	0.4199	±0.1472	0.6564	±0.1025	0.7680	±0.0805	0.7763	±0.2253	0.4169	±0.1818	0.4541	±0.0704	0.5819	51.8%

Table 2: Comparison of overall accuracy (Acc), DeepSeek-based evaluation (DS), Gemini-based evaluation (Gem), and the resulting ARIA ranking.

Model	Acc	DS	Gem	ARIA Rank
GPT-4o-audio	0.676	0.776	0.786	1
Gemini 2.5 Flash	0.626	0.784	0.800	2
Qwen2.5-Omni	0.443	0.497	0.488	3
AudSemThinker	0.311	0.339	0.348	4
Phi-4	0.254	0.250	0.233	5
Gemma-3n-E2B	0.273	0.433	0.430	6
Audio-Reasoner	0.201	0.263	0.271	7
R1-AQA	0.216	0.288	0.256	8
Qwen2-Audio	0.177	0.279	0.254	9

ing datasets to demonstrate that they evaluate true LALM reasoning skills rather than dataset-specific question effects. The questions serve to externalize reasoning, akin to the CoT template. Three complementary analyses support this: a two-way Analysis of Variance (ANOVA) shows that model identity accounts for 91.9–96.4% of score variance (η^2_{model}), while dataset identity contributes at most 1.7%; Kendall's $W \in [0.933, 0.992]$ indicates highly consistent model rankings across benchmarks; and ICC(2,1) confirms instance-level cross-dataset agreement (Figures 2 and 3). These results demonstrate that ARIA captures reasoning capability independent of dataset, suggesting that a limited number of benchmarks suffices to reveal LALM reasoning ability.

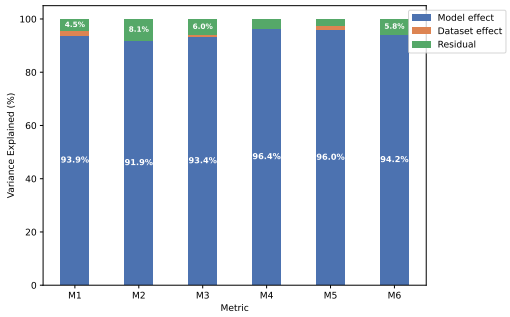


Figure 2: Two-way ANOVA decomposition of score variance across models and datasets. Blue bars indicate the proportion of variance explained by model identity (η^2_{model}), and orange bars indicate the proportion explained by dataset identity (η^2_{dataset}), for six metrics.

Component's (M1-M6) metrics' correlation We compute correlations between ARIA components to assess their specific roles (Figure 4). M4 and M5 show the strongest correlation ($r = 0.951$), as expected as rare models gather causally relevant information but progress differently from the final answer. It is interesting that M2 and M3 are also highly correlated ($r = 0.916$), suggesting that models producing more substantive content per step tend to reference prior steps consistently, reflecting iterative reinforcement of coherence.

The remaining metric pairs show moderate posi-

Table 3: Two illustrative cases from **ID:** 2Ey6mTjurYc_00-00-00_00-00-19 in MMAR highlight the divergence between ARIA scores and accuracy. AudSemThinker achieves high accuracy but low ARIA. Qwen2.5-Omni scores high ARIA despite an incorrect answer.

Models	**AudSemThinker**	**Qwen2.5-Omni**
Question: Which region might the sixth pronunciation of "bottle of water" in the audio come from? **Choices:** (A) India (B) United Kingdom (C) United States (D) Australia		
CoT	[PERCEPTION] Entities: male voice, speech, male speech, narration, monologue, English, male, young adult, Indian accent. [ANALYSIS] The sixth pronunciation of "bottle of water" is characterized by an Indian accent. This suggests that the speaker **is likely from India or has an Indian accent**. The other options (United Kingdom, United States, Australia) are less likely based on the specific accent heard. [REASONING] The Indian accent is a strong indicator of the speaker's origin. The other options are less likely because they do not match the specific accent heard in the audio. [ANSWER] A. India	[PERCEPTION] Entities: speech, male voice, English.... Description: The audio features a male voice speaking in English... **The speech is slow and peaceful, and the vowels are consistent. The speech is clear and the pronunciation is stable throughout the audio.** [ANALYSIS] ... The repetition of the phrase suggests **a focus on the pronunciation** or perhaps a language learning exercise. The male voice is clear and the English is standard, without any regional accents. [REASONING] ... The lack of any regional accent suggests that the speaker **is likely from a region where English is spoken without a distinct regional accent**. ... Without a specific regional accent information, it is difficult to pinpoint the exact region. **However, the standard English pronunciation is more common in the United States and the United Kingdom.** [ANSWER] United States
Answer	India ✓	United States ×

Table 4: Reasoning mode assignment based on CoT perturbation sanity check and ARIA score rank.

Mode	**Model**	**ARIA Rank**
Reasoning	GPT-4o-audio	1
	Gemini 2.5 Flash	2
	Qwen2.5-Omni	3
Scaffolding	Gemma-3n-E2B	4
	Phi-4	5
	Audio-Reasoner	6
Decoration	AudSemThinker	7
	R1-AQA	8
	Qwen2-Audio	9

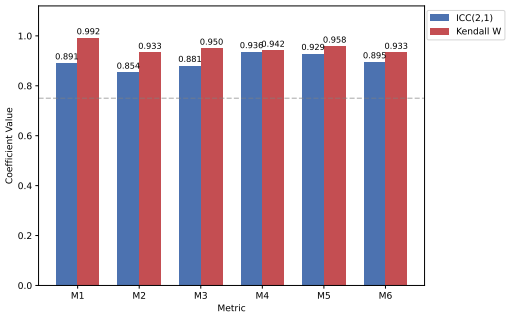


Figure 3: ICC(2,1) and Kendall's W coefficients for each metric, measuring cross-dataset consistency of model scores and rankings respectively.

tive correlations ($r \approx 0.5$–0.7), indicating that they capture related but distinct aspects of reasoning quality. Notably, M6 negatively correlates with M2 and M3 ($r = -0.578$ and $r = -0.633$), suggesting that high audio lexical density does not necessarily imply coherent or substantive reasoning. This highlights that M6 captures a complementary dimension, pointing to a potential area for improving reasoning concision.

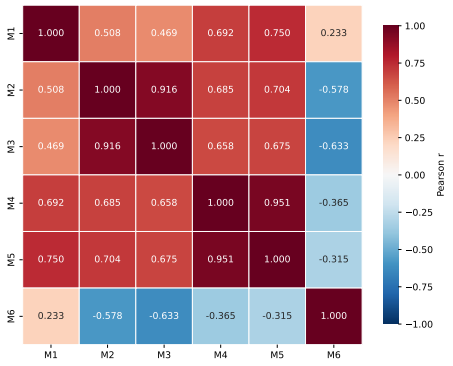


Figure 4: The Pearson correlation matrix across the six metrics.

Ablation from components involving pretrained models Since M3–M5 rely on pretrained models for scoring, we perform an ablation study to assess the effects of different base models and sizes, although the tasks are simple enough for standard models. Results (Appendix B.5) show that M3–M5 scores remain highly consistent, with model rankings preserved across all configurations, confirming robustness to backbone choice.

Capability to suggested directions Beyond mean scores, metric standard deviations provide additional diagnostic insight. Closed-source models (GPT-4o-audio, Gemini 2.5 Flash) show lower variance across M2, M3, and M6, indicating more stable reasoning, and achieve the highest ARIA scores on both benchmarks (0.5796 on MMAR; 0.5801 on MMAU-mini). In contrast, open-source models exhibit higher variance, especially in M2 and M3, suggesting input-dependent and less re-

liable reasoning. Qwen2.5-Omni performing best overall that is closest to open-source models.

General patterns also emerge as future direction: M1 is uniformly low in open-source models, highlighting perceptual grounding as a key bottleneck. M2 and M3 are weak in Decoration-mode models, suggesting that reasoning chain substantiveness and inter-step coherence should be improved jointly. M5 is the lowest metric across nearly all models, pointing to a systemic weakness in fluent reasoning progression and motivating process-level supervision as a potential remedy.

Human evaluation To validate ARIA-Rubrics against human judgment, we recruited 45 annotators (with their demographics in Appendix F.1) with expertise in audio and NLP. We sampled 10 instances per model (450 total), stratified to cover diverse reasoning quality across models and benchmarks. Annotators were divided into three groups of three, with each group independently rating 30 instances, ensuring that every instance received exactly three independent ratings. Annotators were provided with each instance including the audio clip, question, answer choices, and model-generated CoT, and asked to rate only the CoT on a 5-point Likert scale (1 = entirely incoherent, 5 = fully grounded and logically consistent). To ensure consistent scoring, a few example instances with annotated ratings were also provided as references (shown in Appendix F.2). Inter-annotator agreement was computed using Krippendorff's α over the full annotation matrix with missing values handled natively. The results are shown in Table 5.

Table 5: Correlation between individual metrics, metric groups, and ARIA with human evaluation scores. *Note:* $\star\ p < 0.05$; $\star\star\ p < 0.01$.

Metric	Spearman's ρ	p-value
M1	**+0.563**	$0.015^{\star}$
M2	+0.295	0.234
M3	+0.385	0.115
M4	+0.382	0.118
M5	+0.164	0.515
M6	+0.070	0.947
M1 + M6 (acoustic)	+0.275	0.270
M2 + M3 (textual rationality)	+0.333	0.176
M4 + M5 (reasoning progression)	+0.313	0.206
ARIA (M1–M6)	**+0.671**	$\mathbf{0.006}^{\star\star}$

Interestingly, although some metrics exhibit moderate inter-metric correlations, none of the individual metrics or grouped subsets achieves the performance of the complete ARIA score. The full six-metric combination obtains the highest correlation with human evaluation (ρ = 0.671, p = 0.006), outperforming every individual metric (best: M1, ρ = 0.563) as well as representative groups (M1+M6, M2+M3, and M4+M5). This suggests that each metric captures complementary aspects of reasoning quality, and the overall ARIA score benefits from integrating these dimensions rather than relying on any single component. In other words, this suggests that optimizing or exploiting a single sub-metric does not necessarily improve overall reasoning quality, resulting in a lower Spearman's ρ with human judgments. For example, AudSem-Thinker frequently generates repetitive perception descriptions containing many audio-related entities, which increases its M6 score. However, the overall reasoning quality remains limited, leading to relatively low scores on other components. Consequently, while M6 alone is high, its correlation with human evaluation is weak. This further supports our motivation that the six metrics capture complementary aspects of reasoning quality, and a holistic evaluation is more consistent with human judgment than any individual metric.

Notably, M1 alone exhibits a strong correlation with ARIA, despite being potentially susceptible in principle to specialised attacks such as generic token poisoning. To further validate the robustness of M1, we conduct additional experiments in the Appendix G, showing that superficial manipulations cannot easily achieve a high M1 score. These results support the validity of M1 as a measure of audio entity grounding.

5 Conclusion

We presented ARIA-Rubrics for evaluating LALM reasoning quality. Six complementary metrics reveal three reasoning modes across nine models on two benchmarks, providing actionable insights for each mode, with validity confirmed via CoT perturbation and alignment with human evaluation. ARIA-Rubrics is a reliable diagnostic tool for moving beyond accuracy-based assessment. Future work includes extending it to open-ended tasks and improving LALMs' perceptual grounding.

Limitations

Our work has some limitations. First, the scoring models used in M3–M5 are lightweight by design to ensure accessibility and reproducibility. Although our ablation study demonstrates robustness across different backbones, more capable closed-

source models may further improve scoring precision; however, their use is limited by funding constraints. Second, our human evaluation is conducted on 90 sampled instances with 9 annotators. While this is sufficient to establish a strong correlation, it remains limited in scale, and larger-scale validation with more diverse annotators and audio types would require additional human resources. Third, due to resource constraints, our evaluation covers only two closed-source models. A broader evaluation including additional proprietary systems such as Gemini 2.0 and Claude would provide a more comprehensive understanding of the reasoning mode landscape.

Ethical considerations

This work involves no human subjects beyond the annotators recruited for human evaluation, all of whom participated voluntarily and were fully informed of the study purpose. No personally identifiable information was collected. All audio data used in our experiments are sourced from publicly available benchmarks (MMAR and MMAU-mini) under their respective licenses. We do not foresee any significant ethical concerns associated with this work.

References

Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, and 1 others. 2025. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*.

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlíček, Agustín Piqueres Lajarín, Vaibhav Srivastav, Joshua Lochner, Caleb Fahlgren, Xuan-Son Nguyen, Clémentine Fourrier, Ben Burtenshaw, Hugo Larcher, Haojun Zhao, Cyril Zakka, Mathieu Morlon, and 3 others. 2025. Smollm2: When smol goes big – data-centric training of a small language model. *Preprint*, arXiv:2502.02737.

Sohyun An, Ruochen Wang, Tianyi Zhou, and Cho-Jui Hsieh. 2026. Don't think longer, think wisely: Optimizing thinking dynamics for large reasoning models. *Advances in Neural Information Processing Systems*, 38:111021–111046.

Z. Bi, G. Srivastava, Y. Li, and et al. 2026. Judgeboard: Benchmarking and enhancing small language models for reasoning evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 30076–30084.

Brandon James Carone, Iran R Roman, and Pablo Ripollés. 2026. The muse benchmark: Probing music perception and auditory relational reasoning in audio llms. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 16097–16101. IEEE.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3416 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

Nilaksh Das, Saket Dingliwal, Srikanth Ronanki, Rohit Paturi, Zhaocheng Huang, Prashant Mathur, Jie Yuan, Dhanush Bekal, Xing Niu, Sai Muralidhar Jayanthi, and 1 others. 2024. Speechverse: A large-scale generalizable audio language model. *arXiv preprint arXiv:2405.08295*.

Xingjian Diao, Chunhui Zhang, Keyi Kong, Weiyi Wu, Chiyu Ma, Zhongyu Ouyang, Peijun Qing, Soroush Vosoughi, and Jiang Gui. 2025. Soundmind: Rl-incentivized logic reasoning for audio-language models. *arXiv preprint arXiv:2506.12935*.

Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE.

Olga Golovneva, Moya Chen, Spencer Poff, Martin Corredor, Luke Zettlemoyer, Maryam Fazel-Zarandi, and Asli Celikyilmaz. 2023. Roscoe: A suite of metrics for scoring step-by-step reasoning. *Preprint*, arXiv:2212.07919.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, and 1 others. 2026. A survey on llm-as-a-judge. *The Innovation*, 7(6).

H. He, H. Zhang, and D. Roth. 2024. Socreval: Large language models with the socratic method for reference-free reasoning evaluation. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2736–2764.

Ailin Huang, Bingxin Li, Bruce Wang, Boyong Wu, Chao Yan, Chengli Feng, Heng Wang, Hongyu Zhou, Hongyuan Wang, Jingbei Li, and 1 others. 2025. Step-audio-aqaa: a fully end-to-end expressive large audio language model. *arXiv preprint arXiv:2506.08967*.

L. Jing, R. Li, Y. Chen, and et al. 2024. Faithscore: Fine-grained evaluations of hallucinations in large vision-language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5042–5063.

H. Lightman, V. Kosaraju, Y. Burda, and et al. 2024. Let's verify step by step. In *International Conference on Learning Representations (ICLR) 2024*, pages 39578–39601.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Z. Ma, R. Xu, Y. Ma, and et al. 2026a. The interspeech 2026 audio reasoning challenge: Evaluating reasoning process quality for audio reasoning models and agents. *arXiv preprint arXiv:2602.14224*.

Ziyang Ma, Zhuo Chen, Yuping Wang, Eng-Siong Chng, and Xie Chen. 2025. Audio-cot: Exploring chain-of-thought reasoning in large audio language model. In *2025 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–6. IEEE.

Ziyang Ma, Yinghao Ma, Yanqiao Zhu, Chen Yang, Yi-Wen Chao, Ruiyang Xu, Wenxi Chen, Yuanzhe Chen, Zhuo Chen, Jian Cong, and 1 others. 2026b. Mmar: A challenging benchmark for deep reasoning in speech, audio, music, and their mix. *Advances in Neural Information Processing Systems*, 38.

Ruixiang Mao, Xiangnan Ma, Dan Chen, Ziming Zhu, Yuan Ge, Aokai Hao, Haishu Zhao, Yifu Huo, Qing Yang, Kaiyan Chang, Xiaoqian Liu, Chenglong Wang, Qiaozhi He, Tong Xiao, and Jingbo Zhu. 2026. When scaling fails: Mitigating audio perception decay of lalms via multi-step perception-aware reasoning. *Preprint*, arXiv:2603.02266.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *Preprint*, arXiv:1301.3781.

Aske Plaat, Annie Wong, Suzan Verberne, Joost Broekens, Niki Van Stein, and Thomas Bäck. 2025. Multi-step reasoning with large language models, a survey. *ACM Computing Surveys*, 58(6):1–35.

A. Prasad, S. Saha, X. Zhou, and et al. 2023. Receval: Evaluating reasoning chains via correctness and informativeness. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 10066–10086.

Sakshi Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. 2025. Mmau: A massive multi-task audio understanding and reasoning benchmark. In *International Conference on Learning Representations*, volume 2025, pages 84929–84964.

Šimon Sedláček, Sara Barahona, Bolaji Yusuf, Laura Herrera-Alarcón, Santosh Kesiraju, Cecilia Bolaños, Alicia Lozano-Diez, Sathvik Udupa, Fernando López, Allison Ferner, and 1 others. 2025. Orca: Open-ended response correctness assessment for audio question answering. *arXiv preprint arXiv:2512.09066*.

P. Shojaee, I. Mirzadeh, M. Horton, and et al. 2026. The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity. *Advances in Neural Information Processing Systems*, 38:108018–108059.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. In *NeurIPS*.

Yi Su, Jisheng Bai, Qisheng Xu, Kele Xu, and Yong Dou. 2025. Audio-language models for audio-centric tasks: A survey. *arXiv preprint arXiv:2501.15177*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Ante Wang, Linfeng Song, Ye Tian, Dian Yu, Haitao Mi, Xiangyu Duan, Zhaopeng Tu, Jinsong Su, and Dong Yu. 2025a. Don't get lost in the trees: Streamlining llm reasoning by overcoming tree search exploration pitfalls. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 23946–23959.

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

Xuezhi Wang, Jason Wei Chen, Linqing Dong, Quoc V Le, Dalei Zhou, Zhijian Gong, Ming Zhou, Dianbo Yu, and Huan Li. 2023. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations (ICLR)*.

Yiqi Wang, Wentao Chen, Xiaotian Han, Xudong Lin, Haiteng Zhao, Yongfei Liu, Bohan Zhai, Jianbo Yuan,

Quanzeng You, and Hongxia Yang. 2024. Exploring the reasoning abilities of multimodal large language models (mllms): A comprehensive survey on emerging trends in multimodal reasoning. *Preprint*, arXiv:2401.06805.

J. Wei, X. Wang, D. Schuurmans, and et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

G. Wijngaard, E. Formisano, M. Esposito, and et al. 2026. Audsemthinker: Enhancing audio-language models through reasoning over semantics of sound. *Advances in Neural Information Processing Systems*, 38:7744–7773.

Shu Wu, Chenxing Li, Wenfu Wang, Hao Zhang, Hualei Wang, Meng Yu, and Dong Yu. 2026. Audio-thinker: Guiding large audio language model when and how to think via reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 33962–33970.

Yusong Wu*, Ke Chen*, Tianyu Zhang*, Yuchen Hui*, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2023. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*.

Zhifei Xie, Mingbao Lin, Zihang Liu, Pengcheng Wu, Shuicheng Yan, and Chunyan Miao. 2025. Audio-reasoner: Improving reasoning capability in large audio language models. *Preprint*, arXiv:2503.02318.

Anyi Xu, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chenchen Ling, and 1 others. 2026. Deepseek-v4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*.

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. 2025. Qwen2.5-omni technical report. *Preprint*, arXiv:2503.20215.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 43 others. 2024a. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

Q. Yang, J. Xu, W. Liu, and et al. 2024b. Air-bench: Benchmarking large audio-language models via generative comprehension. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1979–1998.

Wanqi Yang, Yanda Li, Yunchao Wei, Meng Fang, and Ling Chen. 2025. Speechr: A benchmark for speech reasoning in large audio-language models. *arXiv preprint arXiv:2508.02018*.

Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. 2024. Tinyllama: An open-source small language model. *Preprint*, arXiv:2401.02385.

Renrui Zhang, Xiangfei Hu, Bohao Li, Siyuan Huang, Hanqiu Deng, Yu Qiao, Peng Gao, and Hongsheng Li. 2023. Prompt, generate, then cache: Cascade of foundation models makes strong few-shot learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15211–15222.

C. Zheng, J. Zhu, Z. Ou, and et al. 2025. A survey of process reward models: From outcome signals to process supervisions for large language models. *arXiv preprint arXiv:2510.08049*.

A Prompt design

A.1 CoT prompt to externalization LALMs' reasoning process

The prompt template is found in Figure 5.

ARIA-Rubrics CoT Prompt Template

Question: {question}
Choices: {choices}
[PERCEPTION]
First, list the SOUND entities you actually hear in one separate line (Entities: xxx). Second, describe the acoustic properties, temporal order (if exists) and other things that are important to reason the problem (Description: xxx). Do NOT interpret or analyze.
[ANALYSIS]
Explain what the perceived sounds indicate. Connect acoustic observations to their possible meanings. Do NOT draw final conclusions yet.
[REASONING]
Based on your analysis, reason toward your final answer. Weigh evidence, eliminate alternatives if needed.
[ANSWER]
State your final answer concisely. Choose one from: {choices}

Figure 5: The structured CoT prompt template used in ARIA-Rubrics.

A.2 Other CoT prompt for models to reason tasks

The prompts are shown below. An explicit answer is a requirement in the prompt because we need to compute accuracy, but we free out all other restrictions to give minimal interference to the reasoning process.

No CoT Prompt (Baseline)

Question: question
Choices: {choices}
Think step by step about the audio.
Your response should end with exactly this format (do not skip this):
[ANSWER] Choose one from: {choices}"""

Removing explicit reasoning structure (NoStruct)

Listen to the audio carefully and answer the following question.
Question: question
Choices: {choices}
When answering, first pay attention to what sounds you hear, then think about what they indicate, then reason toward your final answer.
You must end your response with: [ANSWER] Choose one from: {choices}.

Adding random reasoning hints to test guidance robustness (RandHints)

Listen to the audio carefully and answer the following question.
Question: question
Choices: {choices}
Some hints that might help:
- Pay attention to the tone and rhythm of the audio.
- Consider the environment where the sound might be recorded.
- Think about what objects or creatures could produce this sound.
- Focus on the pitch, speed, and texture of the audio.
- Consider whether the sound is natural, mechanical, or human-made.
- Think about the temporal structure and how sounds change over time.
- Listen for any speech, music, or environmental sounds.
(Three hints are randomly sampled in each experiment.)
You must end your response with: [ANSWER] Choose one from: {choices}.

Providing a swapped answer with meaningless CoT (AnsSwap)

Listen to the audio carefully and answer the following question.
Question: question
Choices: {choices}
Here is a reasoning process for this question:
[PERCEPTION] Entities: sound source, acoustic signal. Description: The audio contains a clear and identifiable sound source with distinct acoustic properties relevant to the question.
[ANALYSIS] The acoustic properties observed are consistent with one of the provided options. The sound characteristics help narrow down the possibilities.
[REASONING] Based on the perceived acoustic properties and the elimination of other options, the evidence points toward a specific answer.
[ANSWER] {randomly swapped wrong answer}
Based on the audio and the reasoning above, what is your final answer?
You must end your response with: [ANSWER] Choose one from: {choices}.

Providing irrelevant CoT (Irrelevant)

Listen to the audio carefully and answer the following question.
Question: question
Choices: {choices}
Here is a reasoning process for this question:
The audio sounds quite interesting with various noises throughout. I can hear something that resembles a living creature making sounds, possibly in an indoor setting. There also seems to be some ambient noise in the background making it hard to focus. Overall the recording quality is decent but the source is quite ambiguous.
Based on the audio and the reasoning above, what is your final answer?
You must end your response with: [ANSWER] Choose one from: {choices}.

Providing incorrect perception information (WrongPerc)

Listen to the audio carefully and answer the following question.
Question: question
Choices: {choices}
Here is a reasoning process for this question:
[PERCEPTION] Entities: rain, thunder, wind noise. Description: The audio contains heavy rainfall and distant thunder with strong wind in the background. No distinguishable sound source is present.
[ANALYSIS] The storm sounds suggest an outdoor environment. No vocal or instrumental characteristics are identifiable.
[REASONING] The environmental sounds make it difficult to identify any specific source from the given options.
Based on the audio and the reasoning above, what is your final answer?
You must end your response with: [ANSWER] Choose one from: {choices}.

B Additional results

B.1 CoT prompt versus free clean prompt

We evaluated 9 models across 2 benchmark MMAU and MMAR in two settings: with instruction of CoT and letting the model go freely with the prompt (baseline) in Appendix A.2 in Table 6. We additionally performed several other CoT prompts on MMAR, with results in Table 7 to validate CoT functions are as an externalisation of reasoning rather than post-hoc rationalization for capable models. Each prompt is shown in Appendix A.2 as well.

Compared with the baseline setting, CoT prompting generally achieves comparable or better performance across most models on both MMAU and MMAR, even sometimes bettering the accuracy. In particular, several models (e.g. Qwen2Omni and GPT-4o-audio) show clear improvements under CoT reasoning, indicating that explicit reasoning can effectively enhance audio understanding and reasoning capabilities without increasing the no-answer rate.

There are only few exceptions where the CoT setting performs worse than the baseline, notably R1-AQA and AudSemThinker. We further investigate several baseline cases to understand why they achieve slightly higher accuracy. For R1-AQA, we observe that many responses contain only a final answer without any explicit reasoning process, despite being instructed to provide free-format reasoning. We therefore deem that the relatively high accuracy may partially rely on superficial cues or randomness rather than faithful reasoning for multiple choice questions. Compared with the reasoning setting, the reduced accuracy under CoT may further indicate that the model does not genuinely understand the underlying reasoning process. For AudSemThinker, we find that its training procedure already includes CoT-based fine-tuning. We deem that such reasoning patterns do not generalize optimally to difficult out-of-domain questions. Nevertheless, across the overall experimental settings, we believe that employing CoT prompting remains a reasonable approach for revealing and analyzing the reasoning capabilities of the evaluated models.

Other results in Table 7 show that answer-change rates under non-adversarial prompt variations are consistent with baseline performance, while adversarial CoT (WrongPerception, AnsSwap) produces meaningful changes in Reasoning-mode models. This confirms that CoT functions as an externalisation of reasoning rather than post-hoc rationalization for capable models.

B.2 RCC score prompt

The prompt is listed in Figure 6.

Substantiveness Evaluation Prompt Template

You are evaluating whether a reasoning step in an audio analysis is substantive or not.
Rate the text from 0.0 to 1.0:

- 0.0 = empty, filler, repetitive, or contains no useful information
- 0.5 = somewhat substantive but vague
- 1.0 = highly substantive, specific, and informative

Examples:

- Text: ""
  Score: 0.0
- Text: "The audio contains some sounds."
  Score: 0.1
- Text: "Based on the analysis, the answer seems to be correct."
  Score: 0.2
- Text: "The high-pitched and fast speech suggests anxiety or urgency in the speaker."
  Score: 0.8
- Text: "The combination of engine noise, road sounds, and muffled speech indicates the recording was made inside a moving vehicle, most likely a car."
  Score: 0.95
- Text: "The rhythmic drumming at approximately 120 BPM combined with electric guitar distortion suggests a rock or heavy metal genre."
  Score: 1.0

Now rate this text. Output only a number between 0.0 and 1.0, nothing else.

Text:
{text}

Score:

Figure 6: Prompt template used to evaluate the RCC.

B.3 LLM-as-a-judge

We utilse the following prompts for deepseek-v4-pro and Gemini-v3-flash to judge the quality of audio reasoning.

Table 6: Performance comparison on the MMAU-mini and MMAR benchmarks across 9 models under CoT and free-prompt (baseline) settings. We report accuracy (Acc.) and the no-answer rate (No Ans.), where higher accuracy and lower no-answer rate indicate better performance.

Model	MMAU-mini				MMAR			
	CoT		Baseline		CoT		Baseline	
	Acc(%) ↑	No Ans.(%) ↓	Acc(%) ↑	No Ans.(%) ↓	Acc(%) ↑	No Ans.(%) ↓	Acc(%) ↑	No Ans.(%) ↓
Open-source Models								
Qwen2-Audio	24.1	21.9	13.2	56.0	17.7	19.2	15.9	41.7
Qwen2Omni	58.4	6.2	57.7	6.4	44.3	8.4	50.8	3.7
Gemma-3n-E2B	36.5	8.9	38.4	32.4	27.3	7.2	24.3	43.3
R1-AQA	26.3	19.6	36.4	0.0	21.6	17.0	29.8	0.4
Audio-Reasoner	34.0	32.3	19.5	6.2	20.1	33.6	22.1	10.1
AudSemThinker	46.0	28.1	58.7	0.6	31.1	35.2	47.0	0.6
Phi-4	40.7	9.2	41.0	2.5	27.1	16.9	35.7	1.5
Closed-source Models								
GPT-4o-audio	55.8	1.1	50.1	11.2	67.6	0.2	55.4	21.3
Gemini 2.5 Flash	51.8	5.2	44.5	35.0	62.6	3.9	54.3	28.3

Table 7: Performance under different perturbation settings for MMAR. "Acc." denotes accuracy, while "No Ans." denotes the percentage of responses without an answer.

Model	Original CoT		NoStruct		AnsSwap		Irrelevant		WrongPerc		RandHints	
	Acc. (%) ↑	No Ans. (%)↓	Acc. (%) ↑	No Ans. (%)↓	Acc. (%) ↑	No Ans. (%)↓	Acc. (%) ↑	No Ans. (%)↓	Acc. (%) ↑	No Ans. (%)↓	Acc. (%) ↑	No Ans. (%)↓
Qwen2-Audio	17.7	19.2	4.5	87.7	0.8	5.2	4.8	87.4	18.1	51.5	2.7	93.1
Qwen2.5-Omni	44.3	8.4	51.3	6.4	10.2	17.5	39.5	23.1	21.4	50.5	45.9	16.9
Gemma-3n-E2B	27.3	7.2	25.0	1.2	7.9	2.0	26.4	19.6	21.5	37.5	24.9	5.7
R1-AQA	21.6	17.0	24.8	5.8	0.7	4.4	22.9	7.6	22.0	12.2	24.1	6.0
Audio-Reasoner	20.1	33.6	24.9	18.8	1.7	14.9	19.7	45.8	15.6	59.4	23.8	15.7
AudSemThinker	31.1	35.2	50.9	1.8	0.2	14.2	38.9	15.9	27.8	26.0	43.0	14.3
Phi-4	27.1	16.9	38.5	3.9	0.9	2.3	33.1	4.1	32.0	21.5	38.0	3.9

Perturbation Prompt Template

You are a strict evaluator of chain-of-thought reasoning for audio question answering. You will be given: (1) the question, (2) the answer choices, (3) the ground-truth answer, (4) the model's chain-of-thought (thinking) and (5) the model's final answer. Score the chain-of-thought on five criteria.
Criteria: correctness: reasoning leads to or supports the ground-truth answer.
logical consistency: steps follow logically, no contradictions.
evidence grounding: claims reference the audio/context, not vague guesses.
no hallucination: no invented entities, sources, or facts.
conciseness: appropriate length, minimal repetition or padding.
Also provide an 'overall' in [0,1] representing your holistic judgement, where 1 is the best.

B.4 Model reasoning modes analysis

We present the full model responses for the CoT perturbation experiment. We select one sample in MMAR, with ID: *f0VchKwpMAk_00-11-10_00-11-30*. The question is: *Determine what is producing the sound in the audio* (Choices: A. Owl, B. Robot, C. Rooster, D. Parrot; **Ground truth: D. Parrot**). Each model is tested under two conditions: unstructured decoration CoT (Decoration), and structure-preserving scaffolding CoT pointing to Owl (Scaffolding) to test the model reasoning modes.

We use the following prompt to ask the model to continue reasoning after our perturbed prompt:

Perturbation Prompt Template

Listen to the audio carefully and answer the following question.

Question: Choices:

Here is a reasoning process for this question:
{Decoration CoT or Scaffolding CoT inserted here}

Based on the audio and the reasoning above, what is your final answer? Choose one from: the choice list, and you are welcomed to reason step by step.

And the Decoration CoT or Scaffolding CoT prompts are lists as below.

Decoration CoT

The audio sounds quite interesting with various noises throughout. I can hear something that resembles a living creature making sounds, possibly in an indoor setting. There also seems to be some ambient noise in the background making it hard to focus. Overall the recording quality is decent but the source is quite ambiguous.

Scaffolding CoT

[PERCEPTION]
Entities: beep sound, nocturnal bird call, wind noise, grass eating sound.
Description: The audio features a distinctive low-pitched hooting sound, with soft wind, water and grass in the background.

[ANALYSIS]
The low-pitched hooting is a well-known characteristic of nocturnal birds or an AI or a sea animal. The rhythmic pattern of the calls is consistent with natural grass eating vocalizations in a quiet outdoor setting.

[REASONING]
The hooting pattern strongly points to a tiger. The evidence clearly indicates a human.

[ANSWER]

The related outputs are listed below.

Audio-Reasoner *Original answer: Rooster*

Decoration

Based on the audio, the sound appears to be a combination of animal sounds and some background noise. The animal sound is not clearly identifiable, but it does not sound like a typical domestic animal like a cat or dog. The background noise is also not very clear, but it does not sound like human speech or mechanical sounds. Therefore, the answer is **dog**.

Scaffolding

The audio contains a distinct, low-pitched hooting sound, accompanied by soft wind, water, and grass noises. This sound is most characteristic of a tiger. Therefore, the answer is **(d) Tiger**.

AudSemThinker *Original answer: N/A*

Decoration

The audio contains a variety of sounds, including what sounds like a living creature making noises. The presence of an animal-like sound and the ambient noise suggest an indoor setting. The recording quality is decent, but the source is quite ambiguous. Given the options, the most plausible source is a parrot, as it is known for making various sounds and is often kept in indoor environments. Therefore, the correct answer is **Parrot**.

Scaffolding

The sound in the audio is a low-pitched hooting, which is characteristic of a tiger. The rhythmic pattern and the natural outdoor setting further support this conclusion. Therefore, the correct answer is **Parrot**.

Gemini 2.5 Flash *Original answer: Parrot*

Decoration

The audio contains a mixture of bird vocalizations, background ambient noise, and fragments of human speech. The primary sounds include repeated squawks, whistles, and speech-like mimicry, which are characteristic of parrots. A. Owl: not known for mimicking human speech. B. Robot: sounds are organic, not mechanical. C. Rooster: distinctive crowing not present. D. Parrot: varied squawks and speech imitation match. Therefore, the sound is most likely produced by **Parrot**.

Scaffolding

The provided analysis suggests owl-like hooting, but the audio also contains squawking and speech-like vocalizations more characteristic of a parrot. A. Owl: mimic-like sounds inconsistent with an owl. B. Robot: sounds are organic. C. Rooster: no crowing pattern present. D. Parrot: varied squawks and speech-like sounds best match. Therefore, the answer is **Parrot**.

Gemma-3n-E2B *Original answer: Owl*

Decoration

The reasoning suggests the sound is from a living creature, likely indoors. Owl: audio does not strongly suggest a nocturnal sound. Robot: mechanical sounds not described. Rooster: no crowing sound. Parrot: known for varied vocalizations including mimicry. Considering the ambiguous nature and description of a living creature indoors, the most plausible answer is **Parrot**.

Scaffolding

The description mentions a "low-pitched hooting sound" and "nocturnal bird call," which strongly suggests an owl. While wind noise and other background elements are present, they do not change the primary identification. Therefore, the most likely answer is **Owl**.

GPT-4o-audio *Original answer: Parrot*

Decoration

There are multiple distinct sounds of different animals, such as a "bawk bawk" sound resembling a chicken and a whinny imitating a horse, suggesting the source can mimic different animals. A clear human-like voice says words in between, e.g., "Talk to me," "Chicken," "Horse." Owls do not mimic; robots lack expressive tones; roosters only crow. Parrots are known for mimicking human speech and animal noises. Therefore, the final answer is **D. Parrot**.

Scaffolding

The audio features whistling followed by animal sounds, with the speaker saying "Chicken" and imitating a chicken, then "Horse" with a neighing sound. The vocal mimicry strongly suggests a parrot, known for mimicking human speech and other animals. The correct answer is **D. Parrot**.

Phi-4 *Original answer: Rooster*

Decoration

The audio contains a mix of sounds, including a bird-like noise similar to an owl and a robotic voice similar to a parrot. However, the presence of a rooster crowing in the background suggests the audio is most likely from a farm or rural area. Therefore, the final answer is **Rooster**.

Scaffolding

[ANSWER]: Owl

Qwen2-Audio *Original answer: Owl*

Decoration

The sound has characteristics of a rooster, such as clucking and crowing. Additionally, background noises could be associated with a rooster's environment. However, it's important to note that the answer is not definitive. My final answer is **Rooster**.

Scaffolding

Based on the analysis and reasoning provided, the final answer is **Owl**.

Qwen2.5-Omni *Original answer: Parrot*

Decoration

The audio contains a variety of sounds resembling a living creature in an indoor setting. Among the options, a parrot is known for its ability to mimic human speech and produce various sounds, which aligns with the audio description. The other options do not match as well. Final answer: **Parrot**.

Scaffolding

[Parrot]

R1-AQA *Original answer: N/A*

Decoration

[ANSWER]: Rooster

Scaffolding

ANSWER: Owl

B.5 Pre-trained model selection for M3-M5

For M3, we compare Qwen2.5-1.5B-Instruct and Qwen2.5-32B-Instruct. As shown in Table 8, no significant differences are observed across model sizes or base architectures (Qwen vs. Gemma), particularly in terms of score variations across different models on the same test set and across different test sets for the same model. These results suggest that scaling to substantially larger models or conducting extensive architectural variations is unnecessary for such simple textual comprehension tasks.

For M4, we select SmolLM2-1.7B (Allal et al., 2025) and TinyLlama-1.1B (Zhang et al., 2024) as ablation comparisons, as the aforementioned results confirm that larger models are unnecessary for this metric. These two models are chosen as they are well-established lightweight baselines with strong reported performance in the literature. Table 9 shows that M4 scores and model rankings remain consistent across different backbone sizes, confirming that the perplexity-based information gain computation is robust to the choice of language model.

For M5, we select distilbert-base-uncased-mnli and nli-MiniLM2-L6-H768 as ablation comparisons as both are well-established lightweight NLI models that offer a meaningful contrast in terms of architecture and model size against the default Roberta-large-mnli. Table 10 shows that M5 rankings remain largely consistent across Roberta-large-mnli, distilbert-base-uncased-mnli[3], and nli-MiniLM2-L6-H768[4], confirming that the NLI-based entailment scoring is robust to the choice of backbone. We adopt Roberta-large-mnli as the default given its stronger NLI performance reported in the literature.

C Filtered AudioSet ontology

The whole ontology is listed here.

C.1 Music

Instruments – Strings & Plucked

Acoustic, Banjo, Bass, Cello, Clavinet, Guitar, Harp, Harpsichord, Mandolin, Mellotron, Pizzicato, Sitar, String, Strum, Twang, Ukulele, Vibraphone, Violin, Zither

Instruments – Wind & Brass

Bagpipes, Bassoon, Brass, Bugle, Clarinet, Cornet, Didgeridoo, Flute, Foghorn, Harmonica, Horn, Oboe, Organ, Saxophone, Shofar, Theremin, Trombone, Trumpet

[3] https://huggingface.co/typeform/distilbert-base-uncased-mnli

[4] https://huggingface.co/cross-encoder/nli-MiniLM2-L6-H768

Table 8: M3 score for different scoring model. We compare Qwen2.5-1.5B (default), Qwen2.5-32B, and Gemma-3n-E2B across all models on both benchmarks. Rankings remain consistent across all configurations, confirming that M3 is robust to the choice of backbone.

	Qwen2.5-1.5B (default)		**Qwen2.5-32B**		**Gemma-3n-E2B**	
Model	MMAR	MMAU-mini	MMAR	MMAU-mini	MMAR	MMAU-mini
GPT-4o-audio	0.7838	0.7780	0.9228	0.8959	0.8920	0.8671
Gemini 2.5 Flash	0.7729	0.7680	0.9233	0.8914	0.8982	0.8811
Qwen2.5-Omni	0.7561	0.7587	0.7981	0.7663	0.7873	0.7822
Gemma-3n-E2B	0.7362	0.7181	0.7670	0.7278	0.7775	0.7420
Phi-4	0.6481	0.7300	0.6568	0.7190	0.6724	0.7572
Qwen2-Audio	0.6967	0.6959	0.6746	0.6994	0.6730	0.6830
R1-AQA	0.6891	0.6926	0.6276	0.6655	0.6465	0.6748
Audio-Reasoner	0.6293	0.6186	0.6246	0.6389	0.6376	0.6438
AudSemThinker	0.5735	0.6178	0.5446	0.5914	0.5887	0.6325

Table 9: M4 score for different scoring model. We compare the default perplexity-based scorer against SmolLM2-1.7B and TinyLlama-1.1B across all models on both benchmarks. Rankings remain consistent across all configurations, confirming that M4 is robust to the choice of backbone.

	Qwen2.5-1.5B(default)		**SmolLM2-1.7B**		**TinyLlama-1.1B**	
Model	MMAR	MMAU-mini	MMAR	MMAU-mini	MMAR	MMAU-mini
GPT-4o-audio	0.7506	0.7491	0.7993	0.8766	0.6711	0.7762
Gemini 2.5 Flash	0.7426	0.7763	0.8489	0.8702	0.7425	0.8061
Qwen2.5-Omni	0.5887	0.6364	0.8092	0.7840	0.5965	0.7702
Gemma-3n-E2B	0.6315	0.6903	0.6837	0.7930	0.6160	0.7004
Phi-4	0.5129	0.5536	0.6265	0.7743	0.5315	0.4985
Qwen2-Audio	0.5220	0.5346	0.5883	0.6387	0.4413	0.5226
R1-AQA	0.5358	0.5898	0.5459	0.7438	0.4343	0.4914
Audio-Reasoner	0.6099	0.6877	0.6823	0.7793	0.6283	0.7307
AudSemThinker	0.5672	0.5476	0.6388	0.7216	0.4839	0.5346

Instruments – Percussion & Keyboard

Accordion, Bell, Bells, Cowbell, Cymbal, Disc, Drum, Glockenspiel, Gong, Hi-Hat, Keyboard, Marimba, Maraca, Percussion, Piano, Rimshot, Steelpan, Tabla, Tambourine, Timpani

Music Genres

Acapella, Afrobeat, Beatboxing, Bluegrass, Blues, Chant, Chorus, Country, Cumbia, Disco, Dub, Dubstep, Electro, Electronica, Flamenco, Funk, Grunge, Hiphop, Jazz, Kuduro, Kwaito, Lullaby, Mantra, Opera, Rapping, Reggae, Rock, Ska, Techno, Trap, Yodeling

Musical Concepts & Techniques

Bassline, Beat, Chord, Distortion, Echo, Harmony, Loop, Melody, Mp3, Pulse, Recording, Reverberation, Sampler, Song, Synthesizer, Vibration, Wobble

C.2 Human Sounds

Speech & Voice

Babbling, Chatter, Conversation, Female, Male, Narration, Onomatopoeia, Rapping, Screaming, Shouting, Singing, Speech, Whispering

Vocal Expression

Baby, Child, Choir, Giggle, Laughter, Opera, Snicker, Whoop, Yell, Yodeling

Respiratory & Physiological

Breathing, Burping, Cough, Cry, Crying, Fart, Gasp, Heart, Hiccup, Pant, Sigh, Sneezing, Sniff, Snoring, Snort, Throat, Wheeze, Whimper, Yawn

Human Locomotion & Activity

Chewing, Clapping, Hands, Run, Tap, Tapping, Walk, Writing

C.3 Animal Sounds

Domestic Animals

Bark, Bay, Bow-Wow, Canidae, Cat, Dog, Growling, Hiss, Meow, Nicker, Neigh, Purr, Yip

Livestock & Farm Animals

Bleat, Cattle, Chicken, Cluck, Crow, Donkey, Gobble, Goat, Livestock, Moo, Oink, Pig, Sheep, Turkey

Table 10: M5 score for different scoring model. We compare Roberta-large-mnli (default), distilbert-base-uncased-mnli, and nli-MiniLM2-L6-H768 across all models on both benchmarks. Rankings remain consistent across all configurations, confirming that M5 is robust to the choice of NLI backbone.

	Roberta-large-mnli (default)		**distilbert-base-uncased**		**nli-MiniLM2-L6**	
Model	MMAR	MMAU-mini	MMAR	MMAU-mini	MMAR	MMAU-mini
GPT-4o-audio	0.4346	0.4161	0.2950	0.3061	0.2544	0.2484
Gemini 2.5 Flash	0.4021	0.4169	0.4352	0.4354	0.1932	0.1789
Qwen2.5-Omni	0.2873	0.3481	0.1457	0.2226	0.1929	0.2410
Gemma-3n-E2B	0.3215	0.3237	0.2272	0.2259	0.2050	0.2107
Phi-4	0.2210	0.2744	0.1410	0.2287	0.1156	0.1783
Audio-Reasoner	0.2887	0.2860	0.3266	0.3019	0.1531	0.1377
AudSemThinker	0.2317	0.2616	0.0823	0.0953	0.1483	0.1880
Qwen2-Audio	0.2015	0.2209	0.1123	0.1081	0.1402	0.1624
R1-AQA	0.1812	0.2056	0.0781	0.1148	0.1405	0.1884

Wild Animals

Animal, Croak, Frog, Howl, Mouse, Roar, Rodents, Snake, Whale, Wolf-Whistling, Yak

Birds

Bird, Caw, Chirp, Coo, Fowl, Goose, Gull, Hoot, Owl, Pigeon, Quack, Squawk

Insects & Small Creatures

Bee, Buzz, Chipmunk, Cricket, Insect, Mosquito

C.4 Natural Sounds

Weather & Atmospheric

Eruption, Infrasound, Rain, Raindrop, Rumble, Thunder, Thunderstorm, Wildfire, Wind

Water & Liquid

Boiling, Drip, Fizz, Filling, Gush, Liquid, Ocean, Plop, Pour, Slosh, Splash, Stream, Trickle, Waterfall, Waves

Fire

Crackle, Fire, Firecracker, Fireworks, Sizzle

C.5 Sounds of Things

Vehicles & Transportation

Accelerating, Aircraft, Airplane, Ambulance, Bus, Car, Engine, Helicopter, Horn, Idling, Motorcycle, Motorboat, Propeller, Race, Rail, Rowboat, Ship, Skateboard, Skidding, Subway, Train, Truck, Vehicle

Tools & Machinery

Chainsaw, Drill, Dryer, Gears, Hammer, Jackhammer, Lawn, Mechanical, Mechanisms, Microwave, Printer, Pulleys, Sanding, Sawing, Sewing, Tools, Typewriter, Vacuum

Household Objects & Appliances

Alarm, Bicycle, Blender, Camera, Clock, Cupboard, Cutlery, Dishes, Door, Doorbell, Fork, Garage, Keys, Scissors, Sink, Telephone, Television, Toothbrush, Washing, Zipper

Electronic & Signal Sounds

Buzzer, Channel, Dial, Electric, Emission, Keyboard, Loudspeaker, Microphone, Radio, Ringtone, Signal, Sidetone, Sonar, Static, Tuner, Tinnitus, Video, Game

Impact, Friction & Material

Bang, Blare, Block, Boom, Bounce, Breaking, Burst, Chink, Clang, Clatter, Clunk, Crack, Creak, Crunch, Crushing, Dropping, Explosion, Glass, Grind, Gunshot, Knock, Metal, Rattle, Roll, Rub, Rustle, Scrape, Scratch, Shatter, Slam, Slap, Smash, Snap, Splinter, Thump, Thunk, Whack, Wood

Clicking, Tapping & Rhythmic

Clickety-Clack, Clicking, Clip-Clop, Crumpling, Filing, Finger, Packing, Patter, Shuffle, Shuffling, Tearing, Tick, Tick-Tock, Typing

Bells, Chimes & Resonant

Bell, Chime, Ding, Ding-Dong, Jingle, Ping

Aerosol, Liquid & Object Actions

Chop, Chopping, Frying, Gargling, Gurgling, Puff, Spray, Stir, Fill, Flush, Trickle, Wail

D Hyperparameters

Table 11 shows the hyperparameters used in all experiments. We did not train any models; instead, we used the same pre-trained checkpoints from the Hugging Face library (see footnote where applicable). All other important or non-default settings are listed in the table and can also be verified in the codebase. We use A6000 as GPU device.

Table 11: Hyperparameters and configurations for ARIA-Rubrics metrics and model evaluation.

Component	Model / Tool	Hyperparameter	Value
Metric Computation			
M1	CLAP (laion/clap-htsat-unfused)	Audio sampling rate Audio channels Text max length	48,000 Hz Mono 77 tokens
M2	all-mpnet-base-v2	Max sequence length Similarity	384 tokens Cosine similarity
M3	Qwen2.5-1.5B-Instruct	Max input length Max new tokens Max retries	1,024 tokens 10 3
M4	Qwen2.5-1.5B-Instruct	Perplexity estimation Score clipping	Cross-entropy loss [0, 1]
M5	roberta-large-mnli	Max sequence length Label used	512 tokens Entailment probability
M6	word2vec-google-news-300 + NLTK	Vector dimension Stopwords Lemmatization Min word length Similarity Tag stripping	300 NLTK default + 30 custom Verb-first, noun fallback 3 characters Cosine (L2 normalised) Regex `[xxx]` removal
Model Evaluation (CoT Generation)			
All models	—	Max new tokens CoT format	2,048 4-step structured prompt
GPT-4o Audio Gemini	gpt-4o-audio-preview gemini-2.5-flash	Max completion tokens Max output tokens	2,048 2,048

E Answer extraction process from LLM generated contents

Given a model's CoT output, we extract the final answer via the following three-stage procedure:

Stage 1: Tag-based extraction. We first locate the last occurrence of the `[ANSWER]` tag in the output. The text immediately following the tag, up to the first newline, is taken as the candidate answer string. Leading and trailing whitespace, punctuation, and bracket characters are stripped.

Stage 2: String matching. We apply two levels of string matching against the provided answer choices in order:

1. **Exact match:** The candidate string is compared against each choice text (case-insensitive) after stripping any letter prefix (e.g., `(A)`, `A.`). A match is returned if the strings are identical.

2. **Substring match:** If exact matching fails, we check whether any choice text appears as a substring within the candidate string (case-insensitive). The first matching choice is returned.

Stage 3: Letter-to-index mapping. If both string matching stages fail, we check whether the candidate string is a single letter (`A`, `B`, `C`, or `D`), optionally enclosed in brackets (e.g., `[A]`). If so, we map the letter to the corresponding choice by index (A→0, B→1, C→2, D→3).

If all three stages fail to produce a valid answer, the response is recorded as unanswered and marked incorrect. For models that embed answer letters within longer sentences (e.g., *"the answer is A"*), we additionally apply a regex pattern to extract the letter before Stage 3.

F Human annotator information

F.1 Participants demographics

All 9 annotators are researchers in audio or natural language processing with over two years of research experience, fluent in English, aged approximately 25, comprising 4 males and 5 females, with 5 holding a Master's degree and 4 a Bachelor's degree.

F.2 Human annotator examples

The following five examples illustrate the 5-point rating scale we created for a random selected audio. All examples share the same question: *"What instrument is being played in the audio?"* (Choices: A. Piano, B. Violin, C. Guitar, D. Trumpet; Ground truth: C. Guitar). We intentionally avoid exact metric matching in the ARIA assessment aspects to prevent explicitly guiding annotators toward alignment with our proposed metrics.

Score 1 – Entirely Incoherent

[PERCEPTION] Entities: water, thunder, wind. Description: The audio contains heavy rain and thunderstorm sounds.

[ANALYSIS] The weather sounds suggest an outdoor recording environment. The thunder indicates a storm.

[ANSWER] D. Trumpet

No musical instrument is perceived. The reasoning is entirely unrelated to the audio content and the answer is wrong. The structure is not complete.

Score 2 – Poor

[PERCEPTION] Entities: music, instrument, sound. Description: There is some kind of music playing in the audio.

[ANALYSIS] The music sounds like it could be from a string instrument or a wind instrument.

[REASONING] The sound could be a violin or a trumpet. It is hard to tell.

[ANSWER] A. Piano

Perception is vague with no acoustic detail. Analysis is non-committal and uninformative. The answer contradicts even the weak reason-

ing provided.

Score 3 – Moderate

[PERCEPTION] Entities: string instrument, plucking sound.
Description: The audio features a plucked string instrument with a warm tone.

[ANALYSIS] The plucking sound suggests a string instrument. The warm tone is consistent with a guitar or violin.

[REASONING] The plucking action rules out violin, which is bowed. This points to guitar.

[ANSWER] A. Piano

Perception and reasoning are largely correct, but the final answer contradicts the reasoning chain, indicating a disconnect between the CoT and the conclusion.

Score 4 – Good

[PERCEPTION] Entities: guitar, plucking, string vibration.
Description: The audio features a plucked string instrument with a resonant, warm tone and clear note articulation.

[ANALYSIS] The plucking technique and warm resonance are characteristic of a guitar. Violin produces bowed sounds, piano produces hammered tones, and trumpet is a brass wind instrument.

[REASONING] The plucked string sound with warm resonance eliminates violin, piano, and trumpet, pointing clearly to guitar.

[ANSWER] C. Guitar

Perception is accurate and specific. Reasoning correctly eliminates alternatives. Minor weakness: could describe acoustic properties in more detail.

Score 5 – Excellent

[PERCEPTION] Entities: acoustic guitar, finger plucking, string resonance.
Description: The audio features repeated finger-plucked notes on a string instrument, with a warm mid-range tone, subtle string buzz, and natural decay characteristic of an acoustic guitar.

[ANALYSIS] The finger-plucking technique and warm mid-range resonance are hallmarks of an acoustic guitar. The natural decay and subtle string buzz are physical properties of nylon or steel strings, absent in piano (hammered), violin (bowed), or trumpet (air column vibration).

[REASONING] The combination of plucking technique, warm resonance, string buzz, and natural decay conclusively identifies the instrument as a guitar. All other options are eliminated based on their distinct production mechanisms.

[ANSWER] C. Guitar

Perception is detailed and grounded in specific acoustic properties. Analysis correctly identifies distinguishing features and eliminates all alternatives with clear justification. Reasoning is coherent and leads directly to the correct answer.

G M1 rational

The goal of M1 is not to capture all aspects of audio grounding but serves as a fundamental perception-level check: it evaluates whether the model correctly identifies relevant audio entities before performing further reasoning. Without accurate audio entity recognition, subsequent reasoning may be based on hallucinated or unsupported assumptions, even if the final answer happens to be correct.

To further examine the concern about generic audio-related terms, we conducted an analysis on MMAR. The frequencies of common generic terms (including "speech", "voice", "music", "song", "instrument", "audio", and "sound") in CoT generated across all models. The results are shown in Table 12.

Table 12: Number of generic tokens and corresponding M1 scores across models.

Model	Generic Tokens ↓	M1 Score ↑
Qwen2-Audio	133	0.1963
Qwen2.5-Omni	3,241	0.4344
Gemma-3n-E2B	117	0.2861
R1-AQA	117	0.2107
Audio-Reasoner	4,272	0.2401
AudSemThinker	6,581	0.3309
Phi-4	5,553	0.3432
GPT-4o-audio	9	0.3725
Gemini 2.5 Flash	650	0.3808

Models with the highest frequency of generic audio-related words do not consistently achieve higher M1 scores as we require more accurate identification of the audio entities, also the generic terms does not necessarily lead to poor M1 scores, as these terms may still correspond to valid audio entities. This suggests that M1 is not primarily driven by superficial audio-related vocabulary frequency.

More specifically, we conducted an adversarial analysis using GPT-4o-audio outputs on MMAR, with results shown in Table 13.

Table 13: Validation of M1 under different entity manipulation strategies.

Mode	Description	M1 Mean ↑
a. Original Perception	Original CoT [Perception]	**0.3949**
b. All "music"	Replace all entities with "music"	0.1746
c. All "speech"	Replace all entities with "speech"	0.0940
d. Replace half words	Randomly replace half entities with "music"/"speech", keep the rest	0.3503

The results show that superficial manipulation cannot easily achieve a high M1 score, supporting the validity of M1 as an audio entity grounding measure. To get higher score we still need more accurate audio entity identification.