WoW-Bench: Evaluating Fine-Grained Acoustic Perception in Audio-Language Models via Marine Mammal Vocalizations

Anonymous ACL submission

Abstract

Large audio language models (LALMs) extend language understanding into the auditory domain, yet their ability to perform low-level listening, such as pitch and duration detection, remains underexplored. However, low-level listening is critical for real-world, out-of-distribution tasks where models must reason about unfamiliar sounds based on fine-grained acoustic cues. To address this gap, we introduce the World-of-Whale benchmark (WoW-Bench) to evaluate low-level auditory perception and cognition using marine mammal vocalizations. WoW-bench is composed of a Perception benchmark for categorizing novel sounds and a Cognition benchmark, inspired by Bloom's taxonomy, to assess the abilities to remember, understand, apply, and analyze sound events. For the Cognition benchmark, we additionally introduce distractor questions to evaluate whether models are truly solving problems through listening rather than relying on other heuristics. Experiments with state-of-the-art LALMs show performance far below human levels, indicating a need for stronger auditory grounding in LALMs.[1]

1 Introduction

Humans naturally perceive, interpret, and reason about sound events in their environment. Motivated by the success of large language models (LLMs) in natural language understanding and reasoning, recent works have developed large audio-lanugage models (LALMs), that apply these capabilities to the auditory domain (Gong et al., 2024; Tang et al., 2024; Ghosh et al., 2024a; Chu et al., 2023; Bai et al., 2023). These models integrate an audio encoder with an LLM to support general audio understanding and instruction following across diverse sound-related tasks. A number of benchmarks have been introduced to evaluate and advance LALMs' auditory understanding and reasoning, including compositional (Ghosh et al., 2024b), deductive (Deshmukh et al., 2025b), and comparative (Deshmukh et al., 2025c) reasoning. More recently, broader benchmarks have been proposed to evaluate a wider spectrum of auditory comprehension and reasoning skills (Sakshi et al., 2025).

However, the perceptual capability that *precedes* the reasoning of LALMs' auditory input remains relatively underexplored, despite its foundational importance for auditory understanding and reasoning. This gap is critical, as perceptual errors often account for a substantial portion of failures in reasoning tasks (Sakshi et al., 2025). Current evaluations of LALMs' perceptual capabilities predominantly rely on audio classification tasks (Gemmeke et al., 2017; Chen et al., 2020; Piczak, 2015; Yang et al., 2024), which map acoustic signals to predefined semantic labels such as "dog barking" or "siren." While these benchmarks assess a model's ability to perform semantic categorization, they offer limited insight into whether models are attending to fine-grained acoustic features, such as pitch and duration. Humans naturally process these acoustic cues prior to deriving semantic meaning (Bregman, 1994).

These low-level acoustic features are essential for forming auditory events and objects (Bregman, 1994; Moore, 2012). Focusing on this perspective, we define *low-level listening* as the perceptual ability to detect and differentiate elementary acoustic attributes prior to semantic interpretation or categorization (Bregman, 1994; Moore, 2012; Nahum et al., 2008). Low-level listening also plays a critical role in understanding and reasoning in unfamiliar scenarios, particularly when combined with knowledge about the acoustic characteristics of novel events. Real world applications such as industrial anomaly detection (Koizumi et al., 2020; Dohi et al., 2022), auditory surveillance systems (Crocco et al., 2016), and bioacoustic moni-
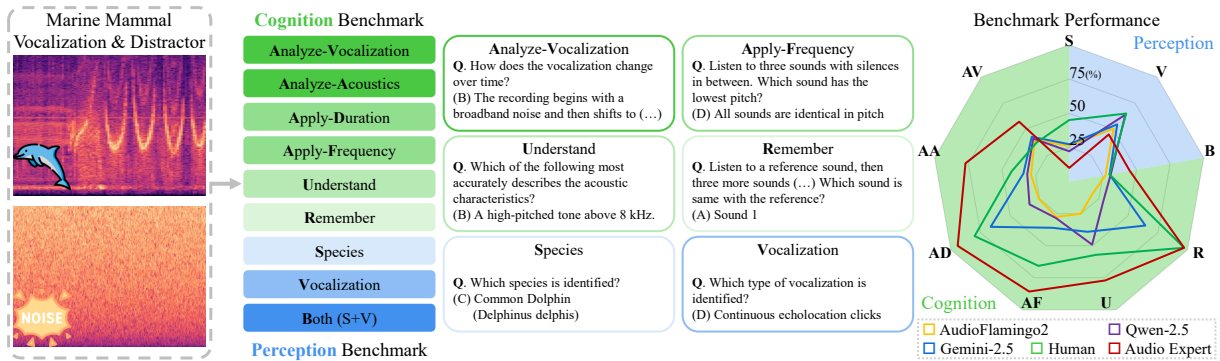

Figure 1: **World-of-Whale** benchmark aims to evaluate low-level listening capabilities of LALMs using marine mammal vocalizations, which are rarely represented in conventional datasets and span a broad acoustic range. LALMs struggle with Cognition questions that humans can reliably solve by using low-level auditory perception.

toring (Bardeli et al., 2010) require strong out-of-distribution (OOD) generalization, as these tasks involve rare and anomalous sounds often absent from training data (Koizumi et al., 2020). Effective models should detect such unheard anomalies by combining knowledge about patterns and low-level acoustic cues such as high-frequency whines.

Recognizing the importance and the relatively overlooked nature of low-level listening, we propose the **World-of-Whale Benchmark (WoW-Bench)** to assess the fine-grained perceptual capabilities of LALMs. Our benchmark focuses on marine mammal vocalizations, which span a broad acoustic range from 20 Hz to over 20 kHz, and exhibit diverse vocalization patterns such as click, whistle, and calls. Moreover, these vocalizations are rarely covered in conventional large-scale audio corpora, *e.g.*, only one out of 527 AudioSet label distributions. By evaluating models in this underrepresented and acoustically rich out-of-domain setting, we aim to provide a more rigorous and fine-grained test of perceptual ability in LALMs.

Our benchmark is structured into two components. First, we assess the perceptual generalization of LALMs by evaluating their ability to categorize sounds into less familiar classes based on low-level listening. Second, we examine their cognitive processing ability, focusing on how well models can interpret and decompose information gained through low-level auditory perception. Inspired by the taxonomy of cognitive hierarchy for learning new concepts (Bloom et al., 1956; Anderson and Krathwohl, 2001), we systematically evaluate models' abilities to *remember*, *understand*, *apply*, and *analyze* auditory information, as illustrated in Figure 1. Additionally, we introduce distractor questions to evaluate whether models are truly solving tasks through listening rather than relying on shallow heuristics or linguistic priors.

Extensive experiments with existing LALMs reveal that both their perceptual generalization and cognitive processing based on low-level listening are limited. As shown in Figure 1, even the best-performing model achieves significantly lower results than humans on the Cognition tasks, highlighting substantial room for improvement in the low-level listening capabilities of LALMs. Moreover, qualitative analysis shows that models tend to adopt a classify-first strategy and infer acoustic properties based on presumed categories rather than listening to them, which can lead to incorrect decisions.

2 Related Work

Large Audio-Language Models. Recent advances in LALMs have substantially improved performance on tasks requiring understanding and reasoning over general sound events. Pengi (Deshmukh et al., 2023) is among the first to unify diverse audio tasks under a single text generation framework by connecting an audio encoder to a decoder-only language model, achieving strong results across a wide range of downstream tasks. Subsequently, a number of LALMs have been introduced (Gong et al., 2024, 2023; Tang et al., 2024; Kong et al., 2024; Chu et al., 2023, 2024; Ghosh et al., 2024a), aligning pretrained audio encoders with large language models and training on large-scale audio-text datasets. These models can follow language instructions and perform a wide range of audio tasks, demonstrating strong performance on both closed-ended tasks (*e.g.*, audio classification) and open-ended tasks (*e.g.*, audio captioning and QA).

More recent efforts focus on enhancing the reasoning capabilities of LALMs by introducing carefully designed training corpora (Ghosh et al.,

2024a, 2025; Deshmukh et al., 2025a; Xie et al., 2025) or incorporating chain-of-thought prompting (Wei et al., 2022; Xie et al., 2025; Ma et al., 2025). While these approaches advance high-level reasoning, comparatively little attention has been paid to low-level auditory perception of the models which serves as a foundation for robust reasoning and understanding (Sakshi et al., 2025). In this work, we address this gap by proposing a new benchmark specifically designed to assess the low-level listening abilities of state-of-the-art LALMs.

Benchmarking Large Audio-Language Models. Following the rapid development of LALMs, several benchmarks have been proposed to assess their capabilities across diverse understanding and reasoning abilities. These benchmarks evaluate compositional reasoning over complex sound events (Ghosh et al., 2024b,a), deductive reasoning via textual entailment conditioned on audio inputs (Deshmukh et al., 2025b), and long-context understanding using long audio segments (Ghosh et al., 2025). Some benchmarks evaluate LALMs across a broad range of domains, including speech, vocal sounds, general audio, and music (Wang et al., 2025; Yang et al., 2024; Sakshi et al., 2025). Among these, AIRBench (Yang et al., 2024) and AudioBench (Wang et al., 2025) primarily focus on audio understanding and instruction-following tasks within the general sound event domain.

However, these benchmarks primarily target reasoning over the perceived sound events, while low-level listening abilities, which is how the model actually listens to and interprets acoustic input remain largely unexplored. Among existing works, ADIFF (Deshmukh et al., 2025c) and MMAU (Sakshi et al., 2025) place relatively greater emphasis on perception. ADIFF addresses comparative reasoning by prompting models to describe the differences between audio clips, while MMAU focuses on information extraction and reasoning across diverse audio types. In contrast, our benchmark is specifically designed to evaluate how LALMs perceive and process novel acoustic events and fine-grained auditory details, thereby assessing their low-level listening capabilities.

Bioacoustics, the study of how animals produce and receive sound, is essential for understanding animal behavior and monitoring ecosystems (Stowell, 2022). From a machine learning perspective, it offers a rich testbed for evaluating auditory perception due to its diverse acoustic environments, wide frequency ranges, and overlapping vocalizations (Rauch et al., 2025). Leveraging these properties, BirdSet (Rauch et al., 2025) introduces a large-scale benchmark for avian vocalizations, highlighting the potential of bioacoustic data for evaluating model robustness and distributional generalization in audio classification tasks. We extend this line of work by focusing on marine mammal vocalizations, which span an exceptionally broad frequency range, from 10 Hz to over 100 kHz (Google Research, 2024), and are underrepresented in standard datasets, even compared to bird sounds. For example, VGGSound (Chen et al., 2020) includes over 10 species-specific labels for birds but only 2 coarse labels for marine mammals. These characteristics make whale sounds a challenging domain for robust evaluation of low-level listening and out-of-distribution generalization in LALMs. Our work differs from the domain adaptation of LALMs to the bioacoustic domain (Robinson et al., 2025) in that we aim to assess the detailed perceptual capabilities of LALMs in novel auditory environments, rather than focusing on the bioacoustics task itself.

3 World-of-Whale Benchmark

3.1 Overview

We introduce the **World-of-Whale Benchmark (WoW-Bench)**, designed to evaluate LALMs on their ability to perceive and cognitively process low-level acoustic details in unfamiliar and acoustically diverse scenarios. The benchmark has two components: (1) The **Perception** benchmark that tests the model's ability to categorize unfamiliar sound events based on low-level listening and their internal knowledge (§3.2), and (2) The **Cognition** benchmark that assesses whether models can cognitively process fine-grained acoustic characteristics and perceived events through low-level listening, as illustrated in Figure 1.(§3.3). We further describe the design of adversarial distractors used in the Cognition benchmark in §3.4, which test whether models are truly listening to solve the questions.

Compared to existing LALM benchmarks, WoW-Bench is the first to explicitly focus on low-level listening and perceptual ability in a controlled OOD setup. As shown in Table 1, prior benchmarks often rely on widely used datasets (Gemmeke et al., 2017; Hershey et al., 2021; Kim et al., 2019; Drossos et al., 2020) or content collected from similar in-the-wild videos (Ju et al., 2024; Islam et al., 2024). Consequently, these benchmarks typically feature similar distribution of sound events. In

Dataset	Target Ability	Low-level Listening	Semantic Source	#Test
CompA (2024b)	Compositional reasoning	✗	Casual, Synthetic	0.6k
CompA-R-*test* (2024a)	Complex reasoning	✗	Casual	1.6k
AIRBench (2024)	Audio understanding & Instruction following	✗	Casual	4.5k*
AudioBench 2025	Audio understanding & Instruction following	✗	Casual	8.9k*
Audio Entailment (2025b)	Deductive reasoning	✗	Casual	5.8k
ADIFF (2025c)	Comparative reasoning	△	Casual	10k
MMAU (2025)	Information extraction & Reasoning	✗	Casual, Synthetic	3.3k*
LongAudioBench (2025)	Long context understanding	✗	Casual, Egocentric	2.4k
WoW-Bench	Perceptual generalization & Cognition based on low-level listening	✓	Marine Mammal (2016)	1.7k

Table 1: Comparison to existing LALM benchmarks. *Low-level Listening* indicates that the dataset evaluates a model's ability to process fine-grained acoustic attributes prior to semantic categorization. *Casual* indicates the dataset corpora covers general sound events like human and music, *e.g.*, Gemmeke et al. (2017); Kim et al. (2019); Drossos et al. (2020); Ju et al. (2024). * denotes the number of general audio-related questions in each test set.

contrast, WoW-Bench utilizes marine mammal vocalizations, which are underrepresented in common corpora. For example, AudioSet contains only one relevant label, VGGSound (Chen et al., 2020) contains two, and both ESC-50 (Piczak, 2015) and FSD50K (Fonseca et al., 2021) contain none. This approach allows for the evaluation of whether models can generalize beyond familiar data and attend to acoustic content, rather than relying solely on training priors.

Identifying suitable OOD audio content is itself a non-trivial challenge, as the pretraining data of many LALMs already encompass a wide range of everyday sounds. We observed that even modest extensions, such as adding other animal vocalizations from the VGGSound test set, can reduce the OOD nature of the task. For instance, when asked to classify animal species, Qwen2-Audio-Instruct's accuracy increased markedly from **28.3%** on WoW-Bench to **76%** on the VGGSound-based variant (e.g., distinguishing Baltimore oriole from pigeon). Additionally, WoW-Bench spans a broad acoustic range from low-frequency sounds of 20 Hz to signals exceeding 20 kHz, covering the entire human auditory spectrum. This makes WoW-Bench particularly well-suited for robust evaluation of low-level listening and to fine-grained acoustic perception.

WoW-Bench consists entirely of multiple-choice questions (MCQs) in light of two observations. While humans may struggle to generate accurate descriptions of unfamiliar sounds, they can often identify the correct option by comparing acoustic details across choices using low-level listening. Moreover, MCQs enable standardized evaluation and are widely adopted for both LLM (Hendrycks et al., 2021; Yue et al., 2024) and LALM benchmarking (Yang et al., 2024; Sakshi et al., 2025).

3.2 Perception Benchmark

The Perception benchmark is designed to judge whether models can classify audio events with unfamiliar labels and acoustic characteristics, based on low-level listening abilities and prior knowledge regarding the problem domain. It measures the perceptual generalization of LALMs to novel domains, where successful classification depends not on memorized patterns but on the model's ability to listen and interpret subtle acoustic features.

(1) The **Species** task requires models to classify each audio clip of a vocalization into the correct species category, such as humpback whale, killer whale, or melon-headed whale. This task is analogous to typical audio event classification and has been a central focus in bioacoustics research. It demands fine-grained auditory discrimination, as vocalization differences between species are often subtle. In our benchmark, we adopt a zero-shot setting, requiring models to map fine-grained perceptual cues to species-level knowledge, if such knowledge exists within the model.

(2) The **Vocalization** task involves models selecting the most appropriate natural language description for a given vocalization. These descriptions refer to vocalization types such as clicks, whistles, or calls, and may also include acoustic characteristics like "high-pitched" or "contains background ship noise." Similarly to the Species task, the Vocalization task requires models to capture low-level acoustic details and map them to unfamiliar labels. However, the vocalization labels are often more intuitive, semantically interpretable, and directly related to the acoustic signal than the name of species. As a result, this task relies more on perceptual matching than on prior knowledge.

(3) The **Both** task requires models to select the op-

tion that correctly describes both the species and the vocalization type for a given audio clip, for example, "Leopard Seal – Long Call." This task presents a more challenging scenario by combining two subtasks and assessing compositional auditory perception. In this setup, models must capture different dimensions of acoustic information simultaneously to perform successfully.

3.3 Cognition Benchmark

Motivated by Bloom's taxonomy (Bloom et al., 1956; Anderson and Krathwohl, 2001), which defines a cognitive hierarchy of learning objectives, we design the Cognition benchmark comprising four subtask types to evaluate different aspects of how LALMs process low-level acoustic details and perceived audio events.

(1) The **Remember** level in the hierarchy involves recognizing and recalling previously encountered information. In this task, the model is given a reference sound, followed by three additional sounds separated by silence. The model must identify which of the subsequent sounds is identical to the reference. This tests the model's ability to recall the reference and recognize similarity based on acoustic characteristics across the segments, without relying on the understanding of sound events.

(2) In Bloom's taxonomy, the **Understand** level refers to the ability to interpret observed information. In our benchmark, we evaluate whether a model comprehends the underlying acoustic properties of a sound by selecting the description that most accurately represents the low-level acoustic features. To reduce ambiguity, each choice includes both a perceptual pitch label and the corresponding frequency range, *e.g.*, "upsweeping high-frequency tonal sound above 8 kHz."

(3) The **Apply** level in the taxonomy denotes the capacity to utilize acquired knowledge in new situations. Accordingly, our Apply task evaluates whether the model can transfer low-level listening abilities to novel contexts. Specifically, the evaluation consists of two parts; in each part, the model compares audio clips based on a core acoustic property, either pitch or duration. In the pitch-based subtask (*Apply–Frequency*), the model receives three clips and identifies the one with either the highest or lowest pitch. In the duration-based subtask (*Apply–Duration*), the model selects the sound with the longest or shortest duration. This task requires the model to utilize low-level perceptual information in a comparative setting, without engaging semantic

understanding of the individual sounds.

(4) Finally, the **Analyze** level in Bloom's taxonomy involves breaking information into parts and examining their relationships. To assess this cognitive skill, we investigate whether the model can analyze transitions within complex acoustic sequences. In particular, we present the model with an audio clip composed of two sounds without any intervening silence and the model must identify and interpret the transition between the sounds. We further divide the task into two subtasks: (i) *Analyze–Acoustics*, which focuses on identifying transitions based on low-level acoustic cues, such as a shift from a low-frequency tone to a broadband pulse; and (ii) *Analyze–Vocalization*, which involves detecting changes in higher-level auditory characteristics like species and vocalization types, such as a transition from a Beluga's squeal to a Common Dolphin's whistle. We carefully construct the answer choices so that attending to only one part of the sequence is insufficient, requiring the model to analyze the shift between segments.

3.4 Distractors and QA Robustness

Sakshi et al. (2025) replace audio inputs with random noise and compare the performance with that on real audio to assess whether models genuinely attend to audio or rely primarily on language priors. However, we believe it is more informative to observe how the model's prediction changes in response to altered audio input, rather than simply verifying the correctness of responses.

Building on this perspective, we develop a more targeted approach by introducing adversarial distractors tailored to each Cognition question type. Each distractor incorporates audio that is qualitatively distinct from those of regular questions for the same type, such that the less expected choice within the context becomes the correct answer, aiming to directly test whether the model is truly solving the task through listening. For example, in the *Apply–Frequency* task, where the model is asked to select the sound with the highest pitch (A. Sound 1, B. Sound 2, C. Sound 3, D. All indistinguishable), we present three identical sounds as input (*i.e.*, the distractor audio). In such case, the model may still choose one of the labeled sounds, as option D can be less expected given the question format. This approach reveals whether the model relies on shallow heuristics or demonstrates genuine listening abilities, thereby providing a more rigorous test of the model's perceptual abilities.

We generate distractor variants by selecting a subset of non-distractor questions and replacing their audio components with distractor audio. We systematically design the distractors by inverting the expected pattern for each question type, thereby encouraging the model to rely on listening to solve the task. For instance, for question types that require comparison across multiple audio samples, we reverse the common expectation that one of the labeled candidate sounds matches the correct answer. Therefore, in the Remember question type, all candidates differ from the reference, making "None of the above" the correct choice. Further details are provided in the Appendix B.2.

3.5 Dataset Curation

All audio samples are sourced from the Watkins Marine Mammal Sound Database (Sayigh et al., 2016). Except for manually constructed Remember questions, all other questions are generated using a large language model and subsequently undergo rigorous human verification and reannotation. For each Perception task, we generate 100 questions. For each Cognition task, including subtasks within Apply and Analyze, we generate 200 questions, along with 50 adversarial distractor questions per task. As a result, the final benchmark comprises **1,777** validated questions, forming a test-only dataset designed for zero-shot evaluation. A detailed description of the curation process and dataset statistics is provided in the Appendix B.3.

4 Experiment

4.1 Setup

Models. We evaluate a range of LALMs capable of processing non-speech sound events, including LTU (Gong et al., 2024), LTU-AS (Gong et al., 2023), SALMONN (Tang et al., 2024), GAMA, and GAMA-IT (Ghosh et al., 2024a), as well as two instruction-following models from the Qwen-Audio series, Qwen-Audio-Chat (Chu et al., 2023) and Qwen2-Audio-Instruct (Chu et al., 2024), on our proposed benchmark. We also consider AudioFlamingo2 (Ghosh et al., 2025), recognized for its strong reasoning and long-context comprehension, and Qwen2.5-Omni (Xu et al., 2025), a multimodal model which exhibits promising performance on audio tasks. For commercial LALMs, we evaluate three models from the Gemini series, specifically Gemini 1.5-Pro (Team et al., 2024), Gemini 2.0-Flash (Google DeepMind, 2024), and Gemini 2.5-Flash-Preview (Comanici et al., 2025), multimodal models known for their strong general audio processing capabilities.

Evaluation Strategy. We report the micro-averaged accuracy across all questions in the benchmark, as well as each task and subtask. To account for varying instruction-following capabilities of different LALMs, we experiment with multiple prompting strategies, *e.g.*, "choose the correct option," "return the answer letter", and report the score of the best-performing prompt for each model. Since LALMs vary in their response formats and do not consistently adhere to instructions regarding formatting, we use GPT-4.1-mini to extract the final answer and determine its correctness. More details of the evaluation prompts and answer extraction are provided in Appendix C.3.

4.2 Results

The main evaluation results are presented in Table 2. Overall, all models perform poorly on the WoW-Bench. Even the best performing commercial model, Gemini-2.5-Flash, remains below 50% accuracy. Most open-source models, except Qwen2.5-Omni, perform at levels similar to random chance. These findings highlight significant limitations in both low-level listening and the associated cognitive processing capabilities of LALMs. The key findings are summarized below:

- Models often perform worse on Cognition than Perception tasks, despite the former requiring no prior knowledge about marine mammals (§4.2.1). In contrast, humans perform significantly better on Cognition tasks (§4.2.2).
- Performance varies substantially across tasks, reflecting strengths and weaknesses in each model's low-level listening capabilities (§4.2.1).
- The performance gap between distractor and non-distractor questions highlights a limitation in auditory grounding (§4.2.3).
- Qualitative examples show that models often adopt a classify-first strategy, which can lead to failure when perceptual grounding is required (§4.2.4).
- Model performance is not significantly affected by acoustic range, even though many LALMs usually works on limited acoustic range (§4.2.5).

6

Model	Params	Perception				Cognition									Total	MMAU
		S	V	B	All	R	U	AF	AD	AA	AV	ND	D	All		
Random		27.3	22.7	22.0	24.0	24.8	25.8	22.0	23.6	28.0	25.3	26.6	23.3	24.9	24.8	–
SALMONN (2024)	13B	24.2	14.4	8.0	15.5	26.8	18.8	32.0	26.4	0.0	3.3	21.9	2.7	18.0	17.6	41.0
LTU (2024)	7B	**28.3**	29.9	24.0	27.4	22.8	30.8	8.4	20.4	22.0	31.95	19.2	<u>36.0</u>	22.6	23.4	22.5
LTU-AS (2023)	7B	24.2	20.6	22.0	22.3	20.0	23.3	18.0	22.0	21.6	26.6	24.3	12.3	21.9	22.0	23.4
GAMA (2024a)	7B	21.2	29.9	24.0	25.0	23.6	22.5	21.6	20.2	20.4	23.2	14.8	51.3	22.2	22.7	41.4
GAMA-IT (2024a)	7B	19.2	24.7	22.0	22.0	19.6	25.0	18.0	18.8	26.8	24.5	12.3	**60.7**	22.1	22.1	43.2
Qwen-Audio-Chat (2023)	8.4B	24.2	36.1	19.0	26.4	30.0	30.4	<u>40.0</u>	24.0	<u>33.2</u>	34.0	<u>38.9</u>	4.67	31.9	31.0	55.3
Qwen2-Audio-Instruct (2024)	8.4B	**28.3**	41.2	19.0	29.4	19.2	30.0	33.2	25.2	27.6	**44.8**	31.0	25.7	29.9	29.8	55.0
Qwen2.5-Omni (2025)	10.7B	22.2	**63.9**	31.0	38.9	28.8	**49.2**	28.4	<u>33.6</u>	31.2	41.1	37.4	26.7	35.3	35.9	**67.9**
AudioFlamingo2 (2025)	3.3B	26.3	50.5	27.0	34.5	19.2	25.0	27.6	25.6	28.4	39.4	29.2	20.7	27.5	28.6	<u>61.6</u>
Gemini 1.5 Pro	–	23.2	35.1	18.0	25.3	28.4	26.7	26.0	30.0	30.0	34.0	35.1	5.7	29.2	28.5	56.8
Gemini 2.0 Flash	–	18.2	46.4	18.0	27.4	<u>43.6</u>	<u>44.6</u>	**42.4**	26.8	<u>33.2</u>	36.5	38.5	35.0	<u>37.8</u>	<u>36.1</u>	56.5
Gemini 2.5 Flash	–	27.3	<u>54.6</u>	30.0	37.2	**64.4**	39.2	36.0	**66.4**	34.0	<u>42.7</u>	54.5	18.3	47.2	45.5	–

Table 2: Evaluation results on the WoW-Bench. Each acronym stands for **S**pecies, **V**ocalization, **B**oth (Perception), **R**emember, **U**nderstand, **A**pply-**F**requency, **A**pply-**D**uration, **A**nalyze-**A**coustics, **A**nalyze-**V**ocalization, **N**on-**D**istractor, **D**istractor (Cognition), respectively. Last two columns report the overall score and their performance on the Sound Test-Mini subset from MMAU (2025) as reference.
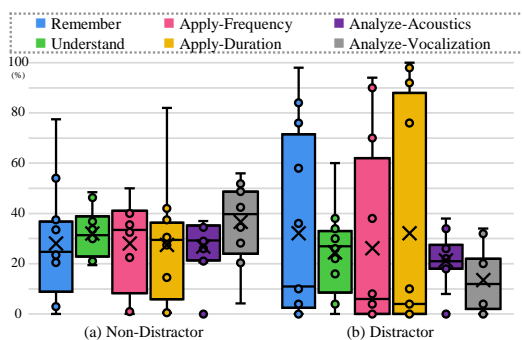

Figure 2: Performance distribution across cognition question types, grouped by the presence of distractors. Each bar represents the interquartile range (Q1 to Q3) of model performance. The three horizontal lines of each bar correspond to the first quartile (Q1), median (Q2), and third quartile (Q3). Circles denote individual model scores, and the × marker indicates the mean.
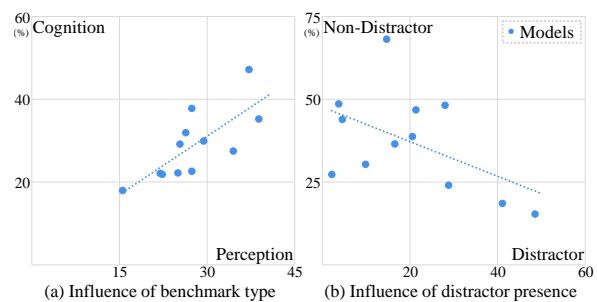

Figure 3: Distribution of models' performance regarding (a) benchmark type and (b) distractor presence, where we observe the Pearson correlation of 0.75 and -0.51, respectively.

4.2.1 How do models perform across different tasks?

One might intuitively assume that the Cognition tasks would be easier than the Perception tasks, as they do not require prior knowledge of marine mammals and can be addressed by simply analyzing the provided audio clip. However, this assumption does not consistently hold true. Aside from the Gemini series models, many models perform worse on the seemingly simpler Cognition tasks. For example, AudioFlamingo2, one of the top-performing models on MMAU, achieves only 19.2% accuracy on the Remember task, lower than its performance on the Species task (26.3%) and even below the random baseline (24.8%). This suggests that audio-grounded cognitive processing remains severely underdeveloped in current LALMs.

In the Perception benchmark, models typically exhibit higher performance on the Vocalization task, where labels such as "whistle" or "clicks" are more intuitively associated with the acoustic signal. The performance on the Species and Both tasks is close to random chance, indicating that models either lack relevant species-level knowledge or fail to capture the nuanced acoustic details necessary for accurate differentiation.

The results on the Cognition benchmark show a more diverse pattern, as illustrated in Figure 2. The performance varies dramatically across both tasks and models; while some models score near zero on certain tasks, others exceed 50% accuracy. This variability highlights the strengths and weaknesses of each model's low-level listening ability. For instance, Gemini-2.5-Flash achieves over 60% accuracy on the Remember and Apply–Duration tasks, both of which involve comparing multiple audio segments. However, its performance drops to 36% on Apply–Frequency, which involves similar comparative analysis, revealing a relative weakness in processing pitch-based information.

Lastly, as shown in Figure 3-(a), models' performance on the Perception and Cognition tasks in WoW-Bench exhibits positive correlation, suggesting that both tasks rely on common low-level listen-

ing capabilities. For comparison, we also include their performance on MMAU Sound Test mini-set in Table 2. Notably, a high score on MMAU does not necessarily translate to strong performance on WoW-Bench. For example, AudioFlamingo2 and SALMONN perform worse than some models on WoW-Bench, despite outperforming them on MMAU. This suggests that WoW-Bench introduces new challenges and evaluation criteria not captured by existing benchmarks.

4.2.2 How do humans perform?

As shown in Figure 1, human and model performance are comparable in the Perception benchmark, as humans often struggle due to limited prior knowledge about marine mammals. However, in the Cognition tasks, humans significantly outperform the models; for example, they achieve 97.1% on the Remember task, while the best Gemini-2.5 model achieves only 57.1 %. These results demonstrate that the questions are well-constructed and reliable; they assess features that humans can readily identify through their low-level listening ability even in an unfamiliar domain, yet remain challenging for current LALMs.

To further evaluate the benchmark's diagnostic value, we include listeners with strong backgrounds in audio signal processing. Expert listeners consistently outperform inexperienced participants Cognition tasks. This indicates that our benchmark tasks effectively capture cognitive processing of fine-grained acoustic properties. Details of the human evaluation are provided in Appendix C.4.

4.2.3 How do models perform on distractors?

As shown in Table 2, models exhibit substantial performance discrepancy between distactor and non-distractor questions. Furthermore, The box plots in Figure 2 illustrate that the performance variation across models for each task type becomes more pronounced for distractor questions. This increased variability suggests that many models rely on linguistic priors or other heuristics rather than consistently attending to low-level acoustic features of the input. Figure 3-(b) additionally shows a negative correlation between model performance on distractor and non-distractor questions. Models that achieve high accuracy on non-distractor questions often struggle with distractor variants, indicating a reliance on non-acoustic cues rather than robust auditory processing.
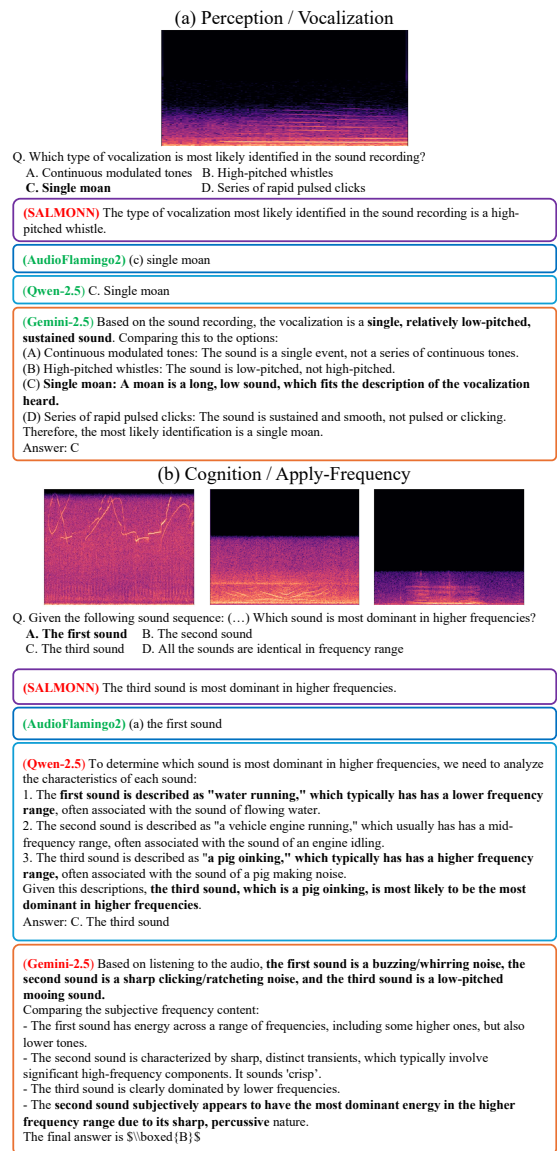

Figure 4: Qualitative examples of model responses to (a) a Perception/Vocalization question and (b) a Cognition/Apply-Frequency question.

4.2.4 How do models respond?

Figure 4 and Figure 5 illustrates LALM responses to WoW-Bench questions. We observe considerable variation in answer formats across models, with many providing direct responses without explicitly analyzing the underlying acoustic features. Notably, in the Perception task, where the model is asked to identify the type of vocalization, Gemini-2.5 attempts to analyze low-level acoustic details and match them to the provided answer choices. In contrast, in the Cognition task involving frequency comparison between three sounds, both, Gemini-2.5 and Qwen2.5-Omni adopt a classification-first approach: they initially assign each sound to a high-level category such as *crisp* or *pig oinking*, and then reason about acoustic attributes based on
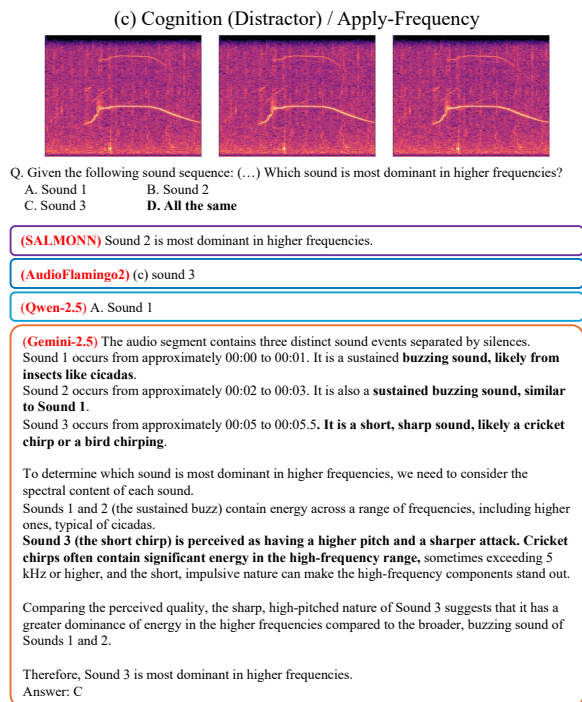
(c) Cognition (Distractor) / Apply-Frequency

Q. Given the following sound sequence: (…) Which sound is most dominant in higher frequencies?
A. Sound 1   B. Sound 2
C. Sound 3   **D. All the same**

(SALMONN) Sound 2 is most dominant in higher frequencies.

(AudioFlamingo2) (c) sound 3

(Qwen-2.5) A. Sound 1

(Gemini-2.5) The audio segment contains three distinct sound events separated by silences.
Sound 1 occurs from approximately 00:00 to 00:01. It is a sustained **buzzing sound, likely from insects like cicadas**.
Sound 2 occurs from approximately 00:02 to 00:03. It is also a **sustained buzzing sound, similar to Sound 1**.
Sound 3 occurs from approximately 00:05 to 00:05.5. **It is a short, sharp sound, likely a cricket chirp or a bird chirping**.

To determine which sound is most dominant in higher frequencies, we need to consider the spectral content of each sound.
Sounds 1 and 2 (the sustained buzz) contain energy across a range of frequencies, including higher ones, typical of cicadas.
Sound 3 (the short chirp) is perceived as having a higher pitch and a sharper attack. Cricket chirps often contain significant energy in the high-frequency range, sometimes exceeding 5 kHz or higher, and the short, impulsive nature can make the high-frequency components stand out.

Comparing the perceived quality, the sharp, high-pitched nature of Sound 3 suggests that it has a greater dominance of energy in the higher frequencies compared to the broader, buzzing sound of Sounds 1 and 2.

Therefore, Sound 3 is most dominant in higher frequencies.
Answer: C

Figure 5: Qualitative example of a model response to (c) a distractor question for Cognition/Apply-Frequency task.

the inferred class, which can often lead to incorrect comparisons. This category-first strategy also contributes to failure in a distractor-type question (Figure 5). For example, when all three sounds are acoustically identical, Gemini-2.5 incorrectly classifies one as different and justifies its answer based on characteristics inferred from the misclassified result, despite the perceptual indistinguishability of all sounds.

These findings suggest that even when the questions require fine-grained acoustic perception, LALMs tend to prioritize semantic classification over perceptual grounding. This tendency highlights a key limitation in their ability to process and reason directly from low-level listening. Additional qualitative analyses are provided in Appendix A.1.

4.2.5 Does acoustic range affect the performance?

By utilizing the wide range of sample rates and durations of samples in WoW-Bench, we analyze how model performance varies with these acoustic properties. As shown in Figure 6, *performance fluctuates with both duration and sample rate, but no sharp increase or drop is observed at any specific threshold.* This suggests that the benchmark fairly reflects the capabilities of open-source models, which typically operate on 16 kHz audio. We
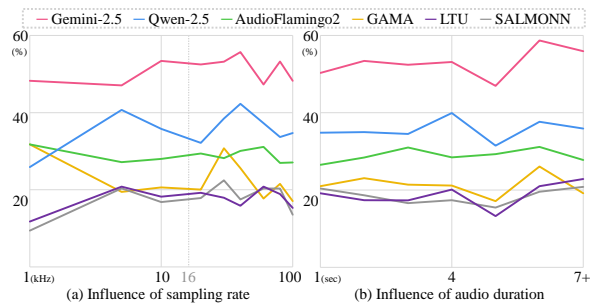

Figure 6: Influence of audio property on accuracy, namely (a) sampling rate and (b) duration.

hypothesize that this phenomenon arises from two factors: (1) while a considerable number of vocalizations extend beyond 10 kHz (*e.g.*, dolphin whistles), a substantial portion of others have their key components within the audible range that most models can process, regardless of the original sample rate of the recordings; and (2) current models do not exhibit significantly improved low-level listening performance even within their nominal frequency range.

These findings suggest two directions for future improvement: (1) enhancing perceptual fidelity within the standard listening range to make better use of accessible acoustic information, and (2) expanding model capacity to process a broader acoustic spectrum, thereby improving generalization to a wider range of real-world audio events.

5   Conclusion

We introduce WoW-Bench, a new benchmark specifically designed to rigorously evaluate the fine-grained, low-level listening, and cognitive capabilities of acoustics-language modeling using marine mammal vocalizations. These results highlight a critical gap in current LALM architectures: despite impressive progress in general audio understanding and instruction following, robust low-level perception remains elusive. Our findings emphasize the necessity for future research to focus on improved auditory grounding and enhanced sensitivity to acoustic detail in order to close the gap between machine and human auditory cognition. WoW-Bench provides a challenging testbed for advancing the next generation of multimodal models.

Limitations

While WoW-Bench provides a rigorous and innovative framework for assessing low-level listening and cognitive processing in large audio-language models (LALMs), several limitations should be acknowledged. First, the benchmark is currently restricted to marine mammal vocalizations. Although this focus is valuable for testing out-of-distribution generalization and robustness, it represents only a narrow segment of the complex diversity found in natural acoustic environments. As a result, our findings may not directly transfer to other domains such as terrestrial bioacoustics, human speech, or complex auditory scenes with overlapping sound sources and background noise.

A further limitation concerns the task format. Our use of multiple-choice questions enables standardized comparisons and the controlled introduction of distractors, but may not fully capture the open-ended reasoning or generative abilities of modern audio-language models. There remains a gap between such discrete evaluation protocols and the continuous, often ambiguous nature of real-world auditory reasoning tasks.

We encourage future research to expand the scope of evaluation, incorporate richer and more interactive assessment paradigms, and explore cross-domain and cross-modal generalization to advance the development of truly robust and perceptually grounded audio-language models.

Ethical Considerations

The creation and use of WoW-Bench raise several ethical considerations. All audio data in the benchmark are sourced from the publicly accessible Watkins Marine Mammal Sound Database, used with explicit permission for research purposes. Care has been taken to ensure that the dataset does not include any personally identifiable information or sensitive content. The potential deployment of audio-language models in ecological monitoring or conservation contexts must consider the ethical implications of automated decision-making, particularly regarding false positives or negatives in species identification, which could impact policy or management actions.

Understanding marine mammal vocalizations is not only of scientific interest but is also crucial for ecological monitoring and marine conservation. Marine mammals play key roles as sentinels of ocean health, and their acoustic behaviors provide unique insights into ecosystem dynamics, species distributions, and the impacts of anthropogenic activities such as shipping and climate change. However, marine bio-acoustics remains an under-resourced field, with limited availability of annotated datasets, research funding, and technological tools compared to terrestrial or human-focused bio-acoustics. Expanding the capabilities of machine listening through responsible audio language models can help bridge this gap, enabling more scalable, continuous, and non-invasive monitoring of marine environments.

References

Lorin W Anderson and David R Krathwohl. 2001. *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives: complete edition*. Addison Wesley Longman, Inc.

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*.

Rolf Bardeli, Daniel Wolff, Frank Kurth, Martina Koch, K-H Tauchert, and K-H Frommolt. 2010. Detecting bird sounds in a complex acoustic environment and application to bioacoustic monitoring. *Pattern Recognition Letters*.

Benjamin S Bloom, Max D Engelhart, Edward J Furst, Walker H Hill, David R Krathwohl, et al. 1956. *Taxonomy of educational objectives: The classification of educational goals. Handbook 1: Cognitive domain*. Longman New York.

Albert S Bregman. 1994. *Auditory scene analysis: The perceptual organization of sound*. MIT press.

Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. 2020. Vggsound: A large-scale audio-visual dataset. In *ICASSP*.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, et al. 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*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

10

Marco Crocco, Marco Cristani, Andrea Trucco, and Vittorio Murino. 2016. Audio surveillance: A systematic review. *ACM Computing Surveys*.

Soham Deshmukh, Satvik Dixit, Rita Singh, and Bhiksha Raj. 2025a. Mellow: a small audio language model for reasoning. *arXiv preprint arXiv:2503.08540*.

Soham Deshmukh, Benjamin Elizalde, Rita Singh, and Huaming Wang. 2023. Pengi: An audio language model for audio tasks. In *NeurIPS*.

Soham Deshmukh, Shuo Han, Hazim Bukhari, Benjamin Elizalde, Hannes Gamper, Rita Singh, and Bhiksha Raj. 2025b. Audio entailment: Assessing deductive reasoning for audio understanding. In *AAAI*.

Soham Deshmukh, Shuo Han, Rita Singh, and Bhiksha Raj. 2025c. Adiff: Explaining audio difference using natural language. In *ICLR*.

Kota Dohi, Keisuke Imoto, Noboru Harada, Daisuke Niizumi, Yuma Koizumi, Tomoya Nishida, Harsh Purohit, Ryo Tanabe, Takashi Endo, Masaaki Yamamoto, and Yohei Kawaguchi. 2022. Description and discussion on dcase 2022 challenge task 2: Unsupervised anomalous sound detection for machine condition monitoring applying domain generalization techniques. In *DCASE2022 Workshop*.

Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. 2020. Clotho: An audio captioning dataset. In *ICASSP*. IEEE.

Eduardo Fonseca, Xavier Favory, Jordi Pons, Frederic Font, and Xavier Serra. 2021. Fsd50k: an open dataset of human-labeled sound events. *IEEE TASLP*.

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 *ICASSP*.

Sreyan Ghosh, Zhifeng Kong, Sonal Kumar, S Sakshi, Jaehyeon Kim, Wei Ping, Rafael Valle, Dinesh Manocha, and Bryan Catanzaro. 2025. Audio flamingo 2: An audio-language model with long-audio understanding and expert reasoning abilities. *arXiv preprint arXiv:2503.03983*.

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

Sreyan Ghosh, Ashish Seth, Sonal Kumar, Utkarsh Tyagi, Chandra Kiran Reddy Evuru, S Ramaneswaran, S Sakshi, Oriol Nieto, Ramani Duraiswami, and Dinesh Manocha. 2024b. Compa: Addressing the gap in compositional reasoning in audio-language models. In *ICLR*.

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

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

Google DeepMind. 2024. Gemini 2.0 flash. https://deepmind.google/technologies/gemini/flash/.

Google Research. 2024. Whistles, songs, boings and biotwangs: Recognizing whale vocalizations with ai. https://research.google/blog/whistles. Accessed: 2025-05-19.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *ICLR*.

Shawn Hershey, Daniel PW Ellis, Eduardo Fonseca, Aren Jansen, Caroline Liu, R Channing Moore, and Manoj Plakal. 2021. The benefit of temporally-strong labels in audio event classification. In *ICASSP*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Md Mohaiminul Islam, Ngan Ho, Xitong Yang, Tushar Nagarajan, Lorenzo Torresani, and Gedas Bertasius. 2024. Video recap: Recursive captioning of hour-long videos. In *CVPR*.

Xuan Ju, Yiming Gao, Zhaoyang Zhang, Ziyang Yuan, Xintao Wang, Ailing Zeng, Yu Xiong, Qiang Xu, and Ying Shan. 2024. Miradata: A large-scale video dataset with long durations and structured captions. In *NeurIPS Datasets and Benchmarks*.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. 2019. Audiocaps: Generating captions for audios in the wild. In *NAACL*.

Yuma Koizumi, Yohei Kawaguchi, Keisuke Imoto, Toshiki Nakamura, Yuki Nikaido, Ryo Tanabe, Harsh Purohit, Kaori Suefusa, Takashi Endo, Masahiro Yasuda, and Noboru Harada. 2020. Description and discussion on dcase2020 challenge task2: Unsupervised anomalous sound detection for machine condition monitoring. In *DCASE2020 Workshop*.

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. In *ICML*.

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. *arXiv preprint arXiv:2501.07246*.

11

Brian CJ Moore. 2012. *An introduction to the psychology of hearing*. Brill.

Mor Nahum, Israel Nelken, and Merav Ahissar. 2008. Low-level information and high-level perception: the case of speech in noise. *PLoS biology*.

OpenAI. 2025. Gpt-4o audio preview. 2025.

Karol J Piczak. 2015. Esc: Dataset for environmental sound classification. In *ACM MM*.

Lukas Rauch, Raphael Schwinger, Moritz Wirth, René Heinrich, Denis Huseljic, Marek Herde, Jonas Lange, Stefan Kahl, Bernhard Sick, Sven Tomforde, and Christoph Scholz. 2025. Birdset: A large-scale dataset for audio classification in avian bioacoustics. In *ICLR*.

David Robinson, Marius Miron, Masato Hagiwara, and Olivier Pietquin. 2025. NatureLM-audio: an audio-language foundation model for bioacoustics. In *ICLR*.

S 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 *ICLR*.

Laela Sayigh, Mary Ann Daher, Julie Allen, Helen Gordon, Katherine Joyce, Claire Stuhlmann, and Peter Tyack. 2016. The watkins marine mammal sound database: an online, freely accessible resource. In *Proceedings of Meetings on Acoustics*.

Dan Stowell. 2022. Computational bioacoustics with deep learning: a review and roadmap. *PeerJ*, 10.

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

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

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

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*.

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

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, et al. 2025. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*.

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

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *CVPR*.

12

A Additioanl Experimental Results

A.1 Additional Qualitative Analysis

We conducted a supplementary error-type analysis to provide more diagnostic insight. Specifically, we examined how the presence of certain words in Gemini 2.5's responses correlates with answer correctness.

The results are shown in Table 3. For *Perception*-type questions, Gemini 2.5 performs relatively well when the response includes acoustic event descriptions such as "hum" or "chirp." However, the model often struggles with identifying specific species, particularly dolphin vocalizations. For *Cognition* questions without distractors, the model performs better when its responses include temporal or presence-related terms such as "longest," "silence," or "occurs." Interestingly, in contrast to Perception tasks, performance degrades when responses include acoustic feature terms such as "repetitive". For *Distractor*-type questions, the trend shifts again; overall performance is lower, but the model shows slightly improved alignment when using global contrastive terms such as "across," "consistent," or "wide." However, it tends to fail in relative comparison, e.g., terms like "longest," "lower," or "third," which it handled relatively well in non-distractor Cognition tasks.

These findings suggest that the reasoning chains used in Cognition tasks differ meaningfully from those used in Perception tasks, and that such reasoning fails to transfer to distractor conditions. This supports our claim that distractor-type questions pose a meaningful challenge and reveal limitations in models' generalization of reasoning strategies. The error-type patterns we observe here are consistent with trends seen in our qualitative analysis

For a more detailed analysis of distractor-type questions, we manually annotated a sample of 44 *Remember*-type items that Gemini 2.5 answered incorrectly. We categorize the errors into four types:

- **No Reasoning (NR)**: The model provides only a bare answer without explanation, e.g., *Answer: B. Sound 2*.
- **Semantic Categorization (S)**: The model selects an answer based on high-level semantic categories rather than acoustic properties, e.g., *"The reference sound is a crying baby. Sound 1 is a mechanical sound. Sound 2 is a mechanical sound. Sound 3 is a crying baby. Sound 3 is the same as the reference sound. Answer: C"*.

- **Low-Level Grounding (L)**: The model reasons based on specific acoustic properties, e.g., *"The reference sound is a steady, high-pitched tone. Sound 1 is a scratching or scrubbing sound. Sound 2 is the same steady, high-pitched tone as the reference sound. Sound 3 is a low-frequency, rhythmic tapping or clicking sound. Sound 2 is the same as the reference sound. Answer: B"*.
- **Direct Comparison (C)**: The model engages in direct comparison with the reference sound, e.g., *"The audio contains a reference sound followed by three other sounds. The reference sound is a short electronic blip. Sound 1 is different from the reference sound. Sound 2 is different from the reference sound. Sound 3 is identical to the reference sound. Therefore, Sound 3 is the same as the reference sound. Answer: C"*.

As shown in Table 4, we observe that when models provide reasoning, semantic categorization is the most frequent strategy. This reflects a category-first behavior, as demonstrated in Figure 5, where the model first classifies the sounds and then infers acoustic characteristics based on those categories. Such behavior suggests that the model is relying on heuristic shortcuts rather than genuine perceptual grounding. While two sounds may belong to the same semantic category, they can still be acoustically distinct. However, the model often treats them as identical solely based on category, failing to capture perceptual differences. This underscores the effectiveness of distractor-type questions in diagnosing whether models are truly attending to low-level acoustic detail.

A.2 Evaluation of Speech-based LALMs

Our goal was to provide a broad and representative evaluation of current LALMs, with a focus on models capable of processing general, non-speech sound events. We intentionally excluded speech-only language models, since our tasks target non-speech auditory processing.

We additionally evaluated GPT-4o-Audio (OpenAI, 2025), but as shown in Table 5, its performance was substantially lower than that of Gemini-2.5-Flash across both the Perception and Cognition benchmarks. In several cases, GPT-4o-Audio produced fallback responses such as "My capabilities include text-based information and analysis, but not audio-based identification," indicating limited

(a) Perception			(b) Non-distractor			(c) Distractor		
Word	Fraction	Accuracy	Word	Fraction	Accuracy	Word	Fraction	Accuracy
hum	25/40	0.6250	longest	152/182	0.8352	wide	29/52	0.5577
chirp	29/54	0.5370	reference	103/134	0.7687	across	20/40	0.5000
modulated	19/36	0.5278	duration	186/265	0.7019	consistent	12/29	0.4138
present	26/55	0.4727	silence	189/278	0.6799	present	11/32	0.3438
continuous	35/75	0.4667	end	91/146	0.6233	modulating	11/32	0.3438
(Average)		0.3716			0.5453			0.1833
(a) Perception			(b) Non-distractor			(c) Distractor		
Word	Fraction	Accuracy	Word	Fraction	Accuracy	Word	Fraction	Accuracy
frequency	14/52	0.2692	broadband	119/334	0.3563	lower	3/35	0.0857
dolphin	35/131	0.2672	harmonic	43/123	0.3496	occur	5/62	0.0806
sperm	14/53	0.2642	repetitive	48/140	0.3429	third	3/61	0.0492
specie	23/103	0.2233	pulse	50/152	0.3289	comparing	4/115	0.0348
common	9/48	0.1875	modulating	45/140	0.3214	longest	1/42	0.0238
(Average)		0.3716			0.5453			0.1833

Table 3: Keyword-based accuracy for the Gemini 2.5 model. Each cell shows the fraction of correct responses when the specified keyword appears in the model's answer, broken down by question type: (a) Perception tasks, (b) Cognition tasks without distractors, and (c) Cognition tasks with distractors

Category	NR	S	L	C	**Total**
Count	16	16	2	10	44

Table 4: Counts of error types in 44 incorrectly answered *Remember*-type distractor questions. R denotes No Reasoning; S denotes Semantic Categorization; L denotes Low-Level Grounding; C denotes Direct Comparison

ability to handle audio input. These observations support our decision to focus on models explicitly designed for general audio processing, rather than those optimized primarily for speech.

B Details of WoW-Bench

B.1 Statistics of WoW-Bench

Our WoW-Bench consists of 1,777 rigorously validated question-answer pairs with audios. Perception benchmark consists of 296 pairs, with 99, 97, and 100 pairs for Species, Vocalization, and Both task, respectively. Each of the six tasks for non-distractor cognition benchmark contains 200 pairs except for Understand and Analyze-Vocalization, for which we filtered out 10 and 9 pairs for their quality. Lastly, regarding distractor-based cognition, we secure a total of 300 pairs, *i.e.*, 50 for each task type. The number of questions for each task type is detailed in Table 6.

B.2 Distractor Design Process

We systematically design the distractors by inverting the expected pattern for each question type, thereby encouraging the model to rely on listening to solve the task. For questions that require comparison across multiple audio samples, we invert the expectation that one of the labeled candidate sounds corresponds to the correct answer. In Remember, all options differ from the reference, making "None of the above" the correct answer. In Apply, all provided sounds are acoustically identical, and the model must correctly select "All are indistinguishable." For tasks focused on understanding the acoustic characteristics of input, we reverse the expectation that the input contains meaningful acoustic content. In Understand, the reference audio is replaced with synthetic noise, requiring the model to avoid hallucinating a semantic interpretation. In Analyze–Acoustics, one segment of a sequence is replaced with noise, and the model must identify the disrupted transition. Lastly, in Analyze–Vocalization, which typically involves detecting species transitions, we eliminate such transitions by concatenating two identical or same-species vocalizations. The model must detect structural redundancy or similarity, rather than blindly assuming a cross-species transition.

To ensure consistency of answer choices across distractor and non-distractor questions, we include the distractor-style options (*e.g.*, "None match" or "All indistinguishable") in Cognition questions. We generate distractor variants by selecting a subset of non-distractor questions and replacing their audio components with distractor audio. For noise-based distractors, we sample from a diverse set of syn-

Model	Perception				Cognition									Total
	S	V	B	All	R	U	AF	AD	AA	AV	ND	D	All	
Gemini-2.5-Flash	27.3	54.6	30.0	37.2	64.4	39.2	36.0	66.4	34.0	42.7	54.5	18.3	47.2	45.5
GPT-4o-Audio	19.2	38.1	19.0	25.3	14.0	26.3	21.2	16.0	24.8	25.3	23.0	14.0	21.2	21.89

Table 5: Evaluation result of Gemini 2.5-Flash and GPT-40-Audio on the WoW-Bench.

Task	Subtask	ND	D	Total
Perception	Species	99	-	99
	Vocalization	97	-	97
	Both	100	-	100
Cognition	Remember	200	50	250
	Understand	190	50	240
	Apply-Frequency	200	50	250
	Apply-Duration	200	50	250
	Analyze-Acoustics	200	50	250
	Analyze-Vocalization	191	50	241
Total		1477	300	1777

Table 6: Number of questions in WoW-Bench by task type. ND denotes non-distractor questions, and D denotes distractor questions.

thetic noise types (*e.g.*, white, pink, brown, blue) that matches the duration and sampling rate of randomly selected real audio clips to ensure consistency in acoustic conditions.

B.3 Details of Data Curation

With the exception of Remember questions, which are manually constructed by selecting candidate sounds, all other questions are generated using a large language model. We provide the model with relevant background information and metadata of the audio clips as input. For questions that require acoustic details not available in the metadata (*i.e.*, Understand and Apply), we employ a vision language model and additionally provide spectrograms of the audio clips. For Analyze-Acoustics, we reuse and adapt acoustic descriptions previously generated and validated from the Understand task to construct transition-based questions. For all question types, we use GPT-4o (`gpt-4o-2024-11-20`) as both the large language model and the vision language model.

For the Perception benchmark, we use metadata such as species names and vocalization descriptions to generate questions and answer choices using GPT-4o (Hurst et al., 2024). For the Cognition benchmark, Remember questions are constructed by manually selecting candidate sounds. Since the Understand and Apply tasks require fine-grained descriptions of acoustic features that cannot be derived from metadata alone, we provide spectrograms to GPT-4o along with detailed guidance for spectrogram interpretation to generate appropriate questions and answer choices. For Analyze–Acoustics, we reuse and adapt acoustic descriptions previously generated and validated from the Understand task to construct transition-based questions. For Analyze–Species, metadata is used to create plausible transitions between species and vocalization types. In both Analyze tasks, GPT-4o is used to generate the final question and answer sets based on selected audio clips.

For each Perception task, we generate 100 questions. For the Cognition tasks, including the subtasks within Apply and Analyze, we generate 200 questions along with 50 adversarial distractor questions per task. We place greater emphasis on the Cognition tasks, as reflected by the higher number of Cognition questions, for two reasons: (1) the Perception questions partly rely on prior knowledge on marine mammals, which current LALMs may not be well-equipped to handle, and (2) the Cognition tasks are more solvable without such prior knowledge, relying primarily on low-level listening. As a result, we treat Cognition scores as a more reliable indicator of perceptual ability in isolation when reporting overall benchmark performance.

All generated questions undergo rigorous human verification and reannotation. Each generated question-answer pair is classified into one of three categories: (1) accept as is, (2) accept with revision, or (3) discard. Similar to the validation pipeline used in MMAU (Sakshi et al., 2025), each question is reviewed by three experienced annotators who cross-check the prompts and QA pairs against the associated metadata and audio clips. If the pair requires revision, annotators evaluate whether it can be rectified with minor edits, such as factual corrections or refinement of uninformative answer choices. If the issue is not fixable, for example, due to high ambiguity or incorrect source data, the pair is discarded. During this process, answer choices are also paraphrased to prevent models from exploiting surface-level lexical cues
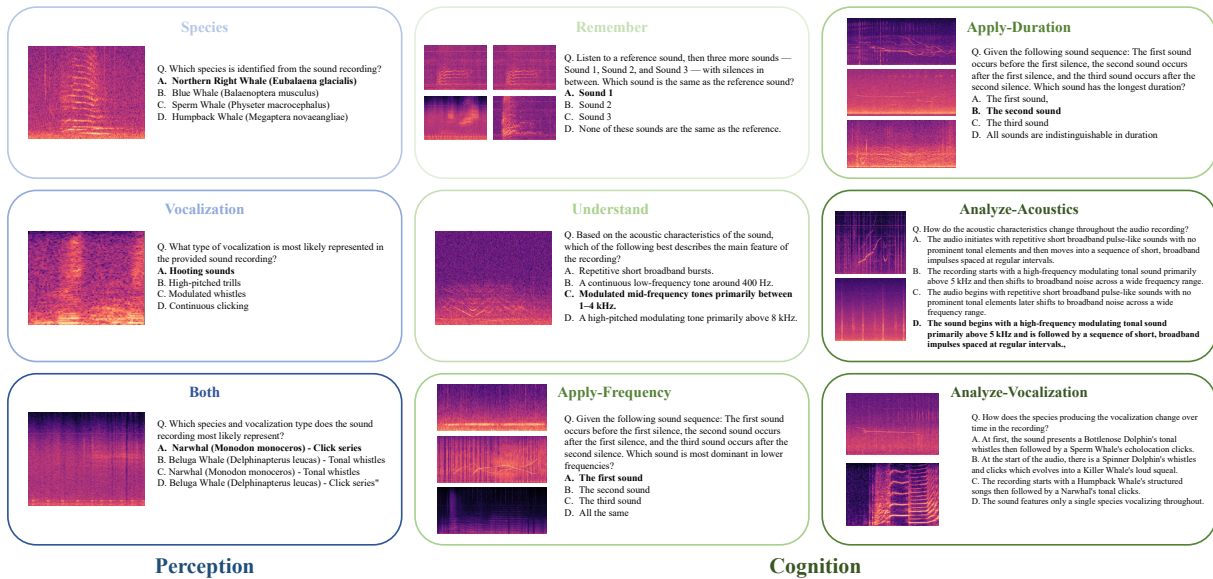
Figure 7: Example questions from each task type in WoW-Bench, shown alongside spectrograms of the corresponding input audio.

B.4 Sample Questions of WoW-Bench

Figure 7 presents representative examples of WoW-Bench questions across all types of tasks.

B.5 Clarification on Design Choices

Q. Why only marine mammal vocalizations? We intentionally designed the benchmark using marine mammal vocalizations to construct a meaningful out-of-distribution (OOD) setup. In familiar sound domains, LALMs often bypass low-level acoustic perception by inferring answers based on high-level semantic associations with known sounds. This behavior undermines the evaluation of genuine low-level listening capabilities, as shown in qualitative examples in Figure 4, and Figure 5 .

Moreover, identifying suitable OOD audio content is non-trivial, as the pre-training data of many LALMs already cover a broad range of everyday sounds, including human speech and urban noise. Even when we modestly extended the dataset by adding other animal vocalizations from the VGGSound test set and asked the model to classify the animal species, performance significantly improved. For instance, Qwen2-Audio-Instruct's accuracy on the species classification task increased from 28.3% in WoW-Bench to 76% in the VGGSound-based set (e.g., Baltimore oriole, pigeon, parrot). This result suggests that broadening the sound categories may compromise the OOD nature of the task and reduce the benchmark's ability to isolate low-level acoustic reasoning. We believe that its focused and controlled design using marine mammal vocalizations enables a more precise evaluation of models' low-level listening capabilities in genuinely OOD settings.

Q. Why use only four levels from Bloom's taxonomy? Bloom's taxonomy consists of six levels of cognitive objectives: Remember, Understand, Apply, Analyze, Evaluate, and Create. Among these, Evaluate and Create are less suitable for modeling low-level listening assessment, since they involve subjective judgment, which are difficult to score objectively, or creative synthesis of new audio. These levels go beyond the scope of low-level listening and the current capabilities of LALMs, which generally cannot generate non-speech audio. Therefore, we focus our benchmark on the first four levels, which are better aligned with the tasks requiring grounded, objective responses based on low-level listening.

Q. Why multiple-choice question format? First, multiple-choice question (MCQ) format is widely adopted in prior LALM benchmarks (Sakshi et al., 2025; Yang et al., 2024) due to its scalability, objective evaluation, and ease of reproducibility. It enables fine-grained control over task structure and difficulty, and facilitates direct comparison across models without requiring prompt-specific evaluation metrics or human scoring. Second, the generative description of marine mammal vocalizations poses a unique challenge. These sounds are difficult to describe even for human annotators, and current models often default to familiar semantic categories instead of grounding their responses in

low-level acoustic properties. For example, when prompted in a manner similar to the Understand task to describe a marine mammal vocalization, the model responded as follows

> **Gemini 2.5-Flash:** Based on the audio provided from 00:00 to 00:04, the acoustic characteristics are:
>
> - **Source:** Avian Vocalizations (specifically bird calls).
> - **Type:** The calls are resonant, distinct, and have a somewhat squawking or crying quality, rather than being simple chirps or melodic songs.
> - **Pattern:** There seems to be a noticeable rhythmic or patterned quality to the sequence of calls.
> - **Number:** It sounds like multiple birds are vocalizing.
> - **Frequency:** The sound primarily occupies the mid-to-high frequency range, typical of bird calls, but with a resonant lower component in the calls.
> - **Dynamics:** Within this short clip, the sound is relatively consistent in volume.
>
> **Reference description:** A series of modulated mid-to-high pitched tones primarily between 2–8 kHz.

The model's output combines high-level semantic content (bird call) with limited references to low-level acoustic features (mid-to-high frequency). However, in a free-form format, such responses are difficult to evaluate reliably, since they vary in specificity and may omit key acoustic details necessary for assessing perceptual grounding.

Third, while our benchmark adopts an MCQ format, it does not preclude open-ended reasoning. Models such as Qwen 2.5-Omni and Gemini 2.5 demonstrate clear intermediate reasoning steps before selecting a final answer as shown in Figure 4, the MCQ structure simply constrains the output space for evaluation purposes, while still allowing models to engage in internal multi-step reasoning, inference, and comparison.

Q. Why evaluate species task in a zero-shot setting? The goal of the species task is to assess whether models can leverage their knowledge about whale species and their vocalizations, potentially acquired from text sources such as Wikipedia, in combination with low-level acoustic cues to make informed predictions. In our results, both models and human participants perform near random chance, which confirms that the task domain effectively represents an OOD scenario. This OOD nature, along with the task's design, highlights a valuable direction for future work: explicitly bridging low-level auditory perception with external knowledge to address the performance gap in such challenging scenarios.

C   Experimental Details

C.1   Models

We enumerate the models and implementations used in all of our experiments as follows:

- **SALMONN** (Tang et al., 2024)[2] (Apache-2.0)
- **LTU** (Gong et al., 2024)[3] (CC BY Attribution 4.0 International)
- **LTU-AS** (Gong et al., 2023)[4] (CC BY Attribution 4.0 International)
- **GAMA** (Ghosh et al., 2024a)[5] (Apache-2.0)
- **Qwen-Audio-Chat** (Chu et al., 2023)[6] (Tongyi Qianwen LICENSE AGREEMENT)
- **Qwen2.5-Omni-7B** (Xu et al., 2025)[7] (Apache-2.0)
- **AudioFlamingo2** (Ghosh et al., 2025)[8] (MIT License)
- **Gemini 1.5 Pro** (Team et al., 2024)[9]
- **Gemini 2.0 Flash** (Google DeepMind, 2024)[10]
- **Gemini 2.5 Flash** (Comanici et al., 2025)[11]

---

[2] https://github.com/bytedance/SALMONN
[3] https://github.com/YuanGongND/ltu
[4] https://github.com/YuanGongND/ltu
[5] https://github.com/Sreyan88/GAMA
[6] https://github.com/QwenLM/Qwen-Audio
[7] https://github.com/QwenLM/Qwen2-Audio
[8] https://github.com/NVIDIA/audio-flamingo
[9] https://arxiv.org/abs/2403.05530
[10] https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-0-flash
[11] https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-5-flash

- **GPT-4o-Audio** (OpenAI, 2025)[12]

To the best of our knowledge, we confirm that our use of aforementioned scientific artifacts is fully compliant of their intended use. All reported results are based on a single run per experiment, where we use one NVIDIA RTX A6000 and 8 CPU cores for running inference with open models.

C.2 Prompts

To ensure fair and standardized evaluation across diverse audio-language models (LALMs), we design task-specific prompts for all inference tasks in WoW-Bench. For multiple-choice questions, the prompt structure clearly presents the question, audio context, and the set of answer options, followed by explicit instructions for answer selection. For example, in the Perception and Cognition benchmarks, a typical prompt is:

> *You will listen to a series of audio recordings. Based on what you hear, choose the most appropriate answer from the options below. Reply with the letter corresponding to your choice.*

Depending on the capabilities and response tendencies of each model, we experiment with minor variations such as: "Please select the correct answer," "Return only the answer letter," or "Explain your reasoning, then provide the answer letter." The best-performing prompt for each model is selected based on preliminary validation.

C.3 Automated Answer Extraction and GPT Evaluation

Given the diverse response formats of LALMs, we employ an automated extraction pipeline to reliably determine model answers. Model outputs are parsed using a lightweight regular expression matcher to identify the final answer letter, regardless of the presence of additional text or reasoning.

For evaluation and quality assurance, we use GPT-4 (or equivalent) to resolve ambiguous cases where the model output does not clearly map to a single answer choice. The evaluation prompt is as follows:

> *Given the following question, options, and the model's response, identify the answer letter (A, B, C, or D) selected by the model. If no clear answer is provided, return "Invalid."*

This automated approach ensures consistent and scalable evaluation, particularly for open-ended or verbose outputs. Ambiguous or invalid responses are excluded from accuracy calculations.

C.4 Human Evaluation

To quantitatively assess human performance on the WoW-Bench tasks, we conducted a targeted human evaluation on a stratified subset of our dataset. From the full benchmark, we randomly sampled 108 question-answer pairs to construct a "mini-test" set, validating that model performance statistics on this subset closely matched those observed on the entire dataset. This approach ensured that the selected subset was both representative and appropriate for reliable human-model comparison.

The 108 questions were divided into two questionnaires, each comprising 54 items, as illustrated in Figure 8 Each questionnaire was independently completed by five participants, resulting in a total of ten unique annotators and 5 times redundancy for every question. In detail, we recruited five inexperienced listeners and five participants with audio-related expertise (e.g., individuals with multiple publications in relevant areas). Detailed results across tasks and participant types are provided in Table 7.

Participants were recruited on a voluntary basis and provided informed consent, with explicit communication regarding the use and anonymization of their response data. Annotators completed the survey remotely, in their environment of choice, and were free to pause or discontinue at any time to mitigate fatigue effects. The user interface was designed for clarity and accessibility: each question included audio playback controls, clearly labeled answer choices, and an optional "Uncertain about the correct answer" option to capture genuine uncertainty and discourage forced guessing.

Detailed instructions were provided at the start of the survey, including recommendations for headphone use and prohibitions on the use of search engines or external reference materials. To minimize potential bias, illustrative audio examples were provided for technical terms such as "*low frequency*," "*high frequency*," and "*broadband pulse*," allowing participants to anchor their perceptual judgments to auditory references rather than textual definitions.

Participants were instructed to base their re-

Category	Gemini 2.5-Flash (%)	Non-Expert (%)	Audio Expert (%)	Human Total (%)
Perception	41.67	**46.67**	33.33	40.00
Cognition: Nondistractor	48.33	62.67	**78.67**	70.67
Cognition: Distractor	25.00	65.00	**88.33**	76.67
Remember	57.14	**97.14**	**97.14**	**97.14**
Understand	42.86	57.14	**77.14**	67.14
Apply–Frequency	21.43	65.71	**85.71**	75.71
Apply–Duration	57.14	80.00	**94.29**	87.14
Analyze–Acoustics	21.43	42.86	**77.14**	60.00
Analyze–Vocalization	50.00	37.14	**57.14**	47.14

Table 7: Detailed human evaluation results across the various task categories.

sponses solely on the presented audio and to avoid making value-based or speculative judgments. Average completion time was less than one hour per participant, and all participants received compensation in accordance with the local legal minimum wage to ensure ethical standards for research participation. This proposed protocol was designed to balance experimental rigor with participant well-being, producing high-quality human baseline data for comparison with model predictions.

Enter User ID

Instruction (click to expand/collapse) ^

- Thank you for participating in our research survey. This survey involves listening to marine mammal vocalizations and selecting the correct answer to multiple-choice questions. Before starting, we strongly recommend wearing earphones or a headset.
- The survey consists of 18 questions in total. Each question includes an audio clip, a question, and four answer choices. Please listen to each audio clip at least once, then select the option you believe is closest to the correct answer.
- If the content is unfamiliar or you remain uncertain of the correct answer even after multiple listens, please check "Uncertain about the correct answer" and select the option that seems most plausible to you.
- You may use dictionaries for less familiar terms, but searching through search engines like Google is **strictly prohibited**.
- Finally, we provide samples below to illustrate concepts that can only be understood by listening to the sounds.
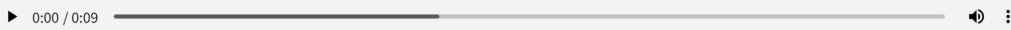- In case of inquiries, please contact:

low frequency   mid frequency   high frequency   broadband pulse

▶ 0:00 / 0:02   🔊 ⋮   ▶ 0:00 / 0:02   🔊 ⋮   ▶ 0:00 / 0:02   🔊 ⋮   ▶ 0:00 / 0:05   🔊 ⋮
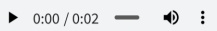
Q6. First, you'll hear a reference sound. Then you'll hear three more sounds — Sound 1, Sound 2, and Sound 3 — each separated by silence. Which sound is identical to the reference sound?

▶ 0:00 / 0:09 ————————————————— 🔊 ⋮

Answer
○
○ A. Sound 1
◉ B. Sound 2
○ C. Sound 3
○ D. No sound is identical to the reference.

☐ Uncertain about the correct answer

Figure 8: Audio-to-text interface of the questionnaire for human evaluation.

20