BEANS-Next and ROOTS: Broadening Audio-Language Capabilities for Bioacoustics

Anonymous authors
Paper under double-blind review

Abstract

Bioacoustics and ethology encompass a wide range of audio understanding tasks, many of which stand to benefit from recent advances in large audio-language models. However, progress in the field has so far been assessed on a narrow set of tasks, primarily centered on label-centric biological category recognition, such as species and call-type classification. In this work, we introduce BEANS-Next, a benchmark grounded in a taxonomy of bioacoustics tasks spanning acoustic perception, biological category recognition, scene understanding, and in-context learning. Using BEANS-Next, we show that existing models exhibit limited performance beyond the task families emphasized by existing evaluations, constraining their usefulness for broader bioacoustic applications. To support progress on this broader task space, we also introduce ROOTS, a large-scale training resource built from expanded curated real-world data and previously underused behavioral and acoustic metadata, supplemented by audio-derived information and scalable synthetic generation where labeling is insufficient. We demonstrate that training on this dataset yields substantial progress across all task groups of BEANS-Next, moving audio-language models closer to their potential as general-purpose assistants for bioacoustics and ethology. To accelerate progress in the field, we open-source our benchmark, dataset, and data pipelines.

1 Introduction

Bioacoustics examines how animals produce, perceive, and use sound. Animal sounds carry information about species identity, behavior, social dynamics, and ecological context, making bioacoustics central to animal behavior research, biodiversity monitoring, conservation, and the study of animal communication (Bradbury & Vehrencamp, 1998; Fischer et al., 2013; Rutz et al., 2023). Machine learning has become a standard tool for analyzing these sounds, with applications including species identification, individual recognition, call-type classification, and behavioral annotation (Stowell, 2022). However, the practical needs of ethologists—researchers who study animal behavior—are broader: researchers often measure acoustic structure (Kim et al., 2026), identify biological categories, count and compare events (Sueur et al., 2008; Holt et al., 2009; Briggs et al., 2012; Marques et al., 2013), reason over temporal patterns (Kershenbaum et al., 2025; Culhane et al., 2025), and adapt analyses to new taxa or study-specific sound classes (Brown et al., 2021; Nieto-Mora et al., 2023; Cauzinille et al., 2024; Kate & Neethirajan, 2025; Kyalo et al., 2025). This breadth makes bioacoustics a compelling domain for large audio-language models (LALMs), which promise a single interface for diverse audio-understanding tasks.

Current LALMs for bioacoustics lack the breadth required to address the full needs of practitioners. Existing benchmarks primarily evaluate classification into a limited subset of biologically relevant categories, especially species or a narrow set of call types (Hagiwara et al., 2023; Rauch et al., 2025b). Many core bioacoustic abilities remain untested: acoustic measurement, call counting, temporal reasoning, call comparison, and adaptation to new taxa or task-specific categories (Sueur et al., 2008; Holt et al., 2009; Briggs et al., 2012). At the same time, standard training datasets provide limited supervision for these abilities. Bioacoustic archives are often weakly labeled and semi-structured, making them difficult to

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

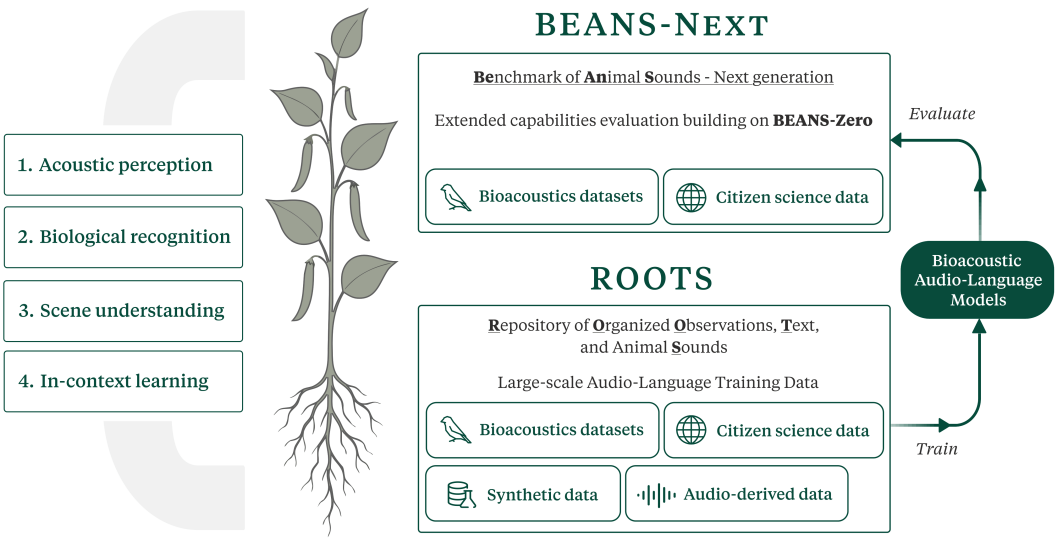

Figure 1: Overview of **BEANS-Next** and **ROOTS**, along with their underlying data sources and the shared four-tier task taxonomy that grounds them.

incorporate into audio-language model training without substantial processing into scientifically meaningful audio-language tasks (Vellinga & Planqué, 2015; Pérez-Granados et al., 2026; Chasmai et al., 2024; Liang et al., 2024). This bottleneck means we currently lack both the evaluation data to measure missing abilities in audio-language models and the training resources needed to make progress.

To address the evaluation side of this bottleneck, we introduce BEANS-Next, a benchmark for measuring broad bioacoustic audio-language abilities needed for ethology research. BEANS-Next is organized around a taxonomy of tasks that reflects both scientific workflows and groups of audio-language model capabilities, spanning acoustic perception, biological category recognition, scene understanding, and in-context learning. It builds on established benchmarks BEANS-Zero (Robinson et al., 2025) and BirdSet (Rauch et al., 2025b), while broadening them across all task groups. Evaluating on BEANS-Next reveals that existing bioacoustic models are narrow, performing competitively on familiar tasks but poorly on broader ones, while general-purpose audio-language models remain poor across most tasks.

To address the training side of this bottleneck, we introduce ROOTS, a large-scale bioacoustic training dataset with 44 million audio-language pairs from 6.5 million distinct audio clips. ROOTS is built from curated bioacoustic archives, acoustic and behavioral metadata that are underused in computational research, scalable synthetic audio pipelines, and audio-derived metadata. We convert semi-structured and unstructured metadata, such as field notes, into instruction-following tasks for species recognition, taxonomy prediction, call-type classification, life-stage prediction, acoustic description, and other biologically meaningful outputs. Where archive labels are weak or incomplete, we use scalable synthetic audio generation pipelines that provide strongly labeled examples for detection, event counting, temporal reasoning, and multi-audio in-context learning. Together, these sources cover the same task groups as BEANS-Next.

We use ROOTS to train bioacoustic audio-language models and evaluate them on BEANS-Next. Starting from a shared NatureLM-audio base model, we fine-tune a separate model on the ROOTS supervision for each tier, isolating the effect of each type of supervision to each ability. Tier-specific training substantially improves performance across all tiers of BEANS-Next, including abilities that are weak or absent in existing models. Because of the aligned taxonomy and training recipe, we show that curated metadata improves biological category recognition, LLM-based text synthesis supports richer description and acoustic perception, and synthetic audio enables scene understanding and in-context capabilities.

Outline and Contributions

(1) We introduce **BEANS-Next**, a benchmark for evaluating bioacoustic audio-language models across **acoustic perception**, **biological category recognition**, **scene understanding**, and **in-context learning**.

(2) We introduce **ROOTS**, a large-scale bioacoustic audio-language training dataset built from curated archives, underused and audio-derived metadata, LLM-based text synthesis, and synthetic audio pipelines.

(3) We characterize a broader **bioacoustic ability space** required by ethology and bioacoustics, and show that existing bioacoustic and general-purpose LALMs cover only a limited portion of this space.

(4) We demonstrate that training on **ROOTS** improves performance across **BEANS-Next**, including capabilities that are weak or absent in existing models, while also improving performance on established bioacoustic tasks.

(5) We release **BEANS-Next** and an evaluation library along with **ROOTS** and its synthetic audio subsetsto support reproducible research on audio-language model evaluation for bioacoustics and ethology.[1]

[1]Available at https://hf.co/iclr2027anon and supplemental material

2 Related Work

With the exception of NatureLM-audio (Robinson et al., 2025), most prior work on LALMs has focused on human speech, music, and general environmental sound. Models such as LTU, SALMONN, Qwen-Audio, Audio Flamingo, and GAMA connect pretrained audio encoders to large language models in order to support instruction following, question answering, captioning, dialogue, and audio reasoning over broad audio inputs (Gong et al., 2024; Tang et al., 2024; Chu et al., 2024; Ghosh et al., 2026; 2024). Complementary benchmarks such as MMAU evaluate increasingly complex audio understanding and reasoning tasks (Sakshi et al., 2025). However, these models and benchmarks are largely centered on human-centric audio distributions, and general audio competence does not directly imply the fine-grained acoustic perception, biological knowledge, or study-specific adaptability required for bioacoustic research.

Bioacoustics has also seen rapid progress in representation learning. Early work showed that large bird-trained embeddings can transfer beyond birds to other taxa and acoustic domains (Ghani et al., 2023; Williams et al., 2025). Self-supervised models such as AVES and animal2vec adapt speech-inspired pretraining objectives to animal vocalizations, improving transfer to downstream bioacoustic classification and detection tasks (Hagiwara, 2023; Schäfer-Zimmermann et al., 2026). More recent models, including Bird-MAE and TweetyBERT, further specialize masked prediction objectives for avian soundscapes and fine-grained birdsong parsing (Rauch et al., 2025a; Vengrovski et al., 2026). While these approaches provide strong acoustic representations, they remain primarily encoder-based and are usually evaluated through fixed-output tasks such as species classification, detection, clustering, or individual identification (Schwinger et al., 2026; Miron et al., 2026).

Language supervision has only recently been introduced into bioacoustic foundation models. BioLingual uses contrastive audio-language pretraining to support zero-shot classification and retrieval from natural language descriptions (Robinson et al., 2024). NatureLM-audio extends this direction to instruction-following audio-language modeling, enabling zero-shot prediction of species, call type, life stage, and other biologically meaningful labels (Robinson et al., 2025). These models demonstrate that language-conditioned bioacoustic models are feasible, but their evaluation remains concentrated on biological category recognition. Broader ethological abilities such as acoustic measurement, event counting, call comparison, temporal reasoning, and adaptation to study-specific sound categories remain underexplored.

Datasets and benchmarks in bioacoustics have similarly focused on recognition problems. BEANS provides a multi-task benchmark for classification and detection across animal sounds (Hagiwara et al., 2023), BIRB evaluates retrieval and generalization in bird vocalizations (Hamer et al., 2023), and BirdSet standardizes large-scale avian soundscape classification under domain shift (Rauch et al., 2025b). Recent audio-language benchmarks have begun to expose the limitations of general LALMs in bioacoustically relevant settings. In particular, WOW-Bench evaluates fine-grained acoustic perception and cognition using marine mammal vocalizations, highlighting the difficulty of low-level listening in unfamiliar acoustic domains (Kim et al., 2026). However, existing benchmarks remain limited in taxonomic scope, task structure, or coverage of the broader abilities needed in ethology. In contrast, BEANS-Next targets a broader bioacoustic ability space across a larger variety of taxa. Appendix Table 10 compares BEANS-Next with existing benchmarks.

Large archives and datasets such as Xeno-canto, iNaturalist Sounds, and the Animal Sound Archive provide substantial taxonomic breadth, but their annotations are often weak, heterogeneous, or semi-structured (Vellinga & Planqué, 2015; Chasmai et al., 2024; Museum für Naturkunde Berlin, 2013). Recent synthetic-data pipelines address part of this limitation by generating strongly labeled examples for sound event detection and few-shot detection (Hoffman et al., 2025; Mahon et al., 2025). These resources support recognition, detection, and data scaling, but are not organized around the broader capabilities needed for general-purpose bioacoustic audio-language models. BEANS-Next and ROOTS address this gap with a benchmark and an aligned training corpus built on the same four-tier taxonomy.

3 Data construction

3.1 A taxonomy of bioacoustic audio-language abilities

While machine learning in bioacoustics has focused primarily on species identification, a broader range of tasks is required by real-world ethology workflows. To structure evaluation and measure progress, we cluster ethological tasks into groups of related tasks we refer to as 'tiers'. These tiers align not only with scientific usage but also with categories of capabilities that have been proposed as relevant to develop and evaluate in LALMs (Ma et al., 2025; Liu et al., 2026; Ghosh et al., 2025). Each has multiple task families, which are comprised of multiple tasks.

Tier 1: Acoustic perception. These tasks measure whether a model can recover the low-level properties of animal sounds, such as fundamental frequency, signal-to-noise ratio (SNR), and acoustic description, without requiring biological category recognition. We instantiate this tier with task families for acoustic-property estimation, acoustic description matching, and acoustic captioning. Scientifically, acoustic measurement is a key part of the workflow of characterizing, interpreting, or comparing acoustic variation (Odom et al., 2021).

Tier 2: Biological category recognition. These tasks ask what biological or ecological categories are present, such as species, broader taxonomic groups, call types, life stage, or behavior. We instantiate this tier with task families for species and taxonomy recognition, taxon presence detection, call-type recognition, life-stage and behavior prediction, and semantic captioning. These are closest to established bioacoustic benchmarks (Hagiwara et al., 2023; Rauch et al., 2025b; Robinson et al., 2025) and support biodiversity monitoring, repertoire analysis, and other analyses of communication.

Tier 3: Scene understanding. These tasks measure whether a model can interpret longer, structured recordings containing multiple vocalizations, species, or individuals. We instantiate this tier with task families for event and species counting, temporal localization, co-occurrence, order reasoning, acoustic comparison across events, species listing, and structured scene description. This setting is especially relevant for passive acoustic monitoring, where recordings often contain overlapping calls, repeated events, multiple individuals or species, and long stretches of background sound. Scientifically, these tasks support questions about not only what occurred, but also when.

Tier 4: In-context learning. These tasks measure whether a model can adapt to a task specified at inference time using text descriptions or audio examples. We instantiate

this tier with task families for audio-example-conditioned classification, few-shot detection, unseen-species recognition, and call-type matching. For example, a researcher may provide examples of an unknown call type, a rare species, or a project-specific sound category and ask the model to identify matching sounds in new recordings. Because bioacoustic data are long-tailed across taxa, tasks, and recording conditions, in-context abilities are critical for learning new tasks at runtime without training a separate classifier for every study.

3.2 Text creation methods

Both BEANS-Next and ROOTS require converting bioacoustic annotations into instruction-following audio-language examples. We use two complementary text creation methods: templated construction and LLM-based construction. These are shared across the benchmark and the dataset. However, BEANS-Next is exclusively grounded on human-annotated metadata, while ROOTS scales supervision by additionally relying on audio-derived metadata, such as acoustic measurements and spectrogram-based descriptions, and synthetic audio.

Templated construction

We construct prompt-response pairs from structured fields such as species identity, taxonomy, locality, call type, life stage, and other available metadata. Each source dataset is mapped into a unified schema, transformed in a task-dependent way, and then converted into multiple task variants using prompt and response templates and algorithmic transformations. The full set of templated tasks and source datasets is provided in Appendix B.1.

LLM-based construction

Templated construction provides reliable supervision when metadata are already structured, but it cannot capture the full range of information available in bioacoustic archives. Many recordings are also accompanied by free-text or semi-structured annotations, such as field notes, that contain scientifically useful information but are difficult to convert into fixed templates without losing nuance. We therefore use an LLM-based pipeline to transform such evidence into natural-language question-answer pairs and descriptions (overview in Appendix B.2.1).

For each task, we write a dedicated prompt specifying the available evidence, the intended task, the output format, constraints on what information may be used, and skip conditions for when the evidence is insufficient or ambiguous. Compared with broad multi-task generation prompts (Ghosh et al., 2026), this single-task design gives finer control over task definitions, their alignment with bioacoustic use cases, and their relative frequency, and narrows each generation problem enough to use a smaller, faster LLM, Gemma 4 26B MoE (Farabet & Lacombe, 2026). Where acoustic supervision is scarce, we also use Gemma's vision capabilities to describe vocalizations characteristics from spectrograms, prompting with human-written examples and restricting descriptions to acoustic properties.

Generated text may contain unsupported claims, inaudible information, references to information that should not be available to the model, or other errors. We therefore filter outputs with regex-based checks for format, keyword, and schema compliance, and with LLM-as-a-judge scores for instruction following with respect to ground truth, hallucination, audibility, and difficulty. One author manually reviewed samples of pre- and post-filtered outputs for each task category as a quality-control step. LLM-based tasks and source datasets are listed in Appendix B.2.

3.3 BEANS-Next benchmark construction

BEANS-Next is constructed as a held-out evaluation suite aligned with the four-tier taxonomy in Section 3.1. Evaluation examples are built from real recordings and human- or dataset-provided annotations. Templated and LLM-based text construction are used only to convert supported annotations into prompts and responses, not to introduce new labels. All BEANS-Next benchmark recordings, and any derivatives of those recordings, are strictly disjoint from ROOTS. For tasks where species identity is central, we additionally follow the held-out species protocol introduced in NatureLM-audio (Robinson et al., 2025): BEANS-Next uses

Table 1: BEANS-Zero classification results (accuracy). Higher is better.

Model	CBI	HBDB	ESC50	Call-type	Lifestage	N-indv
AF-Next	0.006	0.004	0.768	0.078	0.048	0.579
Qwen3-Omni	0.040	0.038	0.797	0.018	0.054	0.531
NatureLM-audio v1.0	0.778	**0.114**	**0.820**	0.870	**0.794**	**0.658**
NatureLM-audio v1.1	**0.808**	0.062	0.360	**0.882**	0.753	0.501

Table 2: BirdSet results (top-1 accuracy). Higher is better.

Model	HSN	NBP	NES	PER	POW	SNE	UHH	AVG
AF-Next	0.000	0.000	0.000	0.000	0.002	0.002	0.000	0.001
Qwen3-Omni	0.003	0.014	0.000	0.000	0.067	0.000	0.000	0.012
NatureLM-audio v1.0	0.273	0.601	0.435	**0.557**	0.891	0.656	0.339	0.536
NatureLM-audio v1.1	**0.325**	**0.644**	**0.446**	0.503	**0.953**	**0.715**	**0.404**	**0.570**

the same held-out species split where applicable, excluding those taxa from the corresponding ROOTS/NatureLM-audio training mixture. We list the held-out taxa in Appendix D.2. Per-tier data sources, task formats, sample counts, and metrics are detailed in Appendix D.

3.4 ROOTS training data construction

ROOTS is an instruction-following training corpus built from curated bioacoustic archives, task-specific annotated datasets, audio-derived metadata, and the synthetic audio pipelines described below. Each source is mapped into a shared schema. Structured fields such as species, taxonomy, call type, life stage, sex, and locality are converted with templates, while free-text notes, acoustic measurements, and spectrogram-derived descriptions are converted with the LLM-based pipeline (Section 3.2). Together, these sources provide supervision across all four tiers of the taxonomy. Source-specific tasks and counts are given in Appendices B and C.5.

Synthetic Audio Pipelines

For most species, bioacoustic data with temporally strong labels are unavailable (Nolasco et al., 2023). In addition, archives used to train bioacoustic foundation models often contain label noise and missing labels (Michaud et al., 2023). To address these gaps, we use heavy augmentation pipelines to synthesize strongly labeled bioacoustic recordings at scale, which are then passed to the text construction pipelines. We summarize the resulting datasets here, with details in Appendix B.3.

We construct two single-audio synthetic datasets. **Strong detection** converts weakly labeled bioacoustic audio into synthetic, strongly labeled recordings with precise timestamps for each species. These recordings are used by the text synthesis pipeline to generate scene understanding tasks. **Detect-diarize** creates synthetic recordings with temporally strong labels and a proxy speaker ID for each call. Starting from BirdMixIT source-separated tracks (Denton et al., 2022), we extract call events and add them to synthetic scenes while preserving their original order and timing and placing them at a consistent amplitude scale.

We also construct synthetic multi-audio datasets. **Few-shot multi-label classification** gives audio examples of up to four target classes and the model must predict which of those classes are present in a query audio clip. **Call-type tasks** use sound clusters to construct synthetic call-type tasks in multiple formats, including multiple choice, binary presence prediction, and same–different prediction between two sounds. Positive examples are sampled from the same species and acoustic cluster according to our processing pipeline, while negatives are sampled from random clusters outside the target cluster.

Table 3: BEANS-Next Tier 1 (acoustic perception) results. Captioning uses CIDEr (×100); description matching and SNR MCQ use accuracy (%); SNR regression and mean f_0 use MAE in dB and Hz, respectively. Higher is better unless marked ↓. Small grey percentages give the parse rate (share of answers with a parseable number); MAE is computed over parsed answers only, and grey MAEs rest on fewer than half of the answers and are not bolded.

	Acoustic description				SNR		Mean f_0 MAE (Hz) ↓	
Model	Caption	Desc. MCQ	Crow		MCQ	MAE (dB) ↓	Seen	Held-out
AF-Next	0.5	50.5	**48.0**		36.7	31.0 2%	5706 14%	268 2%
Qwen3-Omni	0.0	43.6	47.0		25.2	15.4 95%	3749 >99%	208 100%
NatureLM-audio v1.0	0.3	28.4	33.5		34.6	22.6 92%	4086 >99%	208 100%
NatureLM-audio v1.1	0.0	28.7	30.0		26.7	19.1 52%	4334 84%	360 91%
NatureLM-audio T1	**2.8**	**97.9**	27.5		**60.8**	**10.4** 100%	**253** >99%	**125** 100%

4 Experimental Setup

4.1 Models

We evaluate the bioacoustic audio-language model NatureLM-audio v1.0 (Robinson et al., 2025) and two strong open general-purpose LALMs, Audio Flamingo Next (AF-Next) (Ghosh et al., 2026) and Qwen3-Omni (Xu et al., 2025). We focus on generative LALMs because most BEANS-Next tasks require open-ended or multi-audio answers that encoder-only and contrastive models cannot produce; proprietary API models were excluded due to high failure rates and cost (Appendix E). We additionally train NatureLM-audio v1.1 as the base model for our ROOTS experiments; it retains the v1.0 architecture but adds expanded training data and support for longer and multiple audio inputs. All models are queried through the same inference, parsing, and scoring pipeline.

4.2 Training

NatureLM-audio v1.1 is trained on an expanded version of the NatureLM-audio v1.0 semantic task mixture, using curated metadata and primarily templated supervision. From this checkpoint, we fine-tune a separate model on the additional ROOTS supervision for each tier (NatureLM-audio T1–T4), with the audio encoder and Q-former unfrozen and the LLM fine-tuned with LoRA (Hu et al., 2022) (Appendix C.2). Holding architecture and initialization fixed isolates the effect of the training data: because the base model has seen primarily templated semantic supervision, comparisons against it measure the contribution of LLM-based text synthesis, synthetic audio, and audio-derived metadata. We do not train a single joint model in this work; joint multi-tier training, possible interference between tiers, and architectural changes are left to future work.

5 Results

On the established benchmarks (Tables 1 and 2), NatureLM-audio models are far stronger than general-purpose LALMs at species-level classification. On BirdSet, Audio Flamingo Next and Qwen3-Omni score near zero, mostly answering 'Bird' or 'None'. NatureLM-audio v1.1 improves over v1.0 on CBI, call type, and BirdSet, but regresses on ESC-50, HumBugDB, life stage, and individual counting.

5.1 Tier 1: Acoustic Perception

Table 3 shows that ROOTS Tier 1 training adds signal-level capabilities that are weak in the NatureLM-audio v1.1 base model, improving acoustic descriptions, SNR estimation, and mean f_0 prediction. The improvement on unseen-species f_0 estimation is particularly important because it suggests that the model is not only exploiting taxon-specific priors, but has learned a more transferable mapping from audio to quantitative acoustic attributes. The

Table 4: BEANS-Next Tier 2 (biological category recognition) results. Presence and behavior use accuracy (%), fixed-vocabulary call type uses macro-F1 (%) averaged over the five labels, and captioning uses CIDEr (×100). Higher is better.

	Taxon presence				Call-type presence					
Model	Bird	Mamm.	Amph.	Insect	Alarm	Flight	Beg.	Call type	Behav.	Caption
AF-Next	71.2	67.1	81.2	75.5	56.2	46.5	54.9	44.4	56.5	0.1
Qwen3-Omni	68.4	65.0	85.7	73.4	66.8	57.0	65.7	25.6	56.7	0.1
NatureLM-audio v1.0	55.7	63.4	68.0	56.0	50.3	51.7	50.2	48.2	39.9	**1.7**
NatureLM-audio v1.1	54.7	65.2	53.2	59.3	58.8	71.5	62.5	39.5	33.7	1.1
NatureLM-audio T2	**91.5**	**87.9**	**96.4**	**93.4**	**69.9**	**82.2**	**74.6**	**59.0**	**87.1**	1.2

Table 5: BEANS-Next Tier 3 (scene understanding) results. Classification and temporal tasks use top-1 accuracy (%); counting tasks use mean absolute error (MAE); the frequency-range task (Range) uses coverage-aware species-band IoU (%, mean over ground-truth species, missed species scored 0); summary and captioning use corpus CIDEr (×100). `S`: species, `V`: vocalization, `#`: number of. Higher is better unless marked ↓. Small grey percentages give the parse rate (share of answers with a parseable number); MAE is computed over parsed answers only, and grey MAEs rest on fewer than half of the answers and are not bolded.

	Counting MAE ↓			Temporal		Scene / open		
Model	`#S`	`#V`	`#V/S`	Referring	Overlap	Range	Sum.	Caption
AF-Next	1.64 97%	4.26 >99%	2.15 53%	33.5	47.0	0.0	0.0	0.0
Qwen3-Omni	**1.38** 97%	3.93 >99%	1.94 88%	25.0	46.3	0.0	0.0	0.0
NatureLM-audio v1.0	1.54 86%	4.09 100%	1.94 2%	28.0	59.0	0.0	0.1	0.0
NatureLM-audio v1.1	1.50 77%	4.57 93%	1.49 13%	30.0	47.1	0.0	2.4	0.1
NatureLM-audio T3	1.78 100%	**3.81** 100%	**1.72** >99%	**48.0**	**61.7**	**11.3**	**5.0**	**1.6**

	Species ID by characteristic									
	Open-ended					MCQ				
Model	Order	High	Low	Long	Freq.	Order	High	Low	Long	Freq.
AF-Next	0.4	3.8	0.0	5.3	0.0	31.4	35.0	50.3	35.6	30.5
Qwen3-Omni	1.0	0.5	0.2	0.4	0.6	36.1	49.5	48.8	43.5	34.8
NatureLM-audio v1.0	34.7	15.9	13.8	14.2	27.0	37.7	43.4	36.2	43.5	40.7
NatureLM-audio v1.1	30.6	33.3	25.9	**37.3**	36.0	29.2	31.6	32.0	29.5	30.1
NatureLM-audio T3	**41.9**	**37.0**	**32.0**	33.2	**38.7**	**47.2**	**61.8**	**58.1**	**48.6**	**51.1**

main exception is crow vocalization description, where the general-purpose LALMs remain better; this likely points to residual sensitivity to domain and annotation format, rather than uniform transfer across all acoustic-description settings.

5.2 Tier 2: Biological Category Recognition

Table 4 shows that ROOTS Tier 2 training also improves biological category recognition. The gains across presence, call-type, and behavior prediction suggest that mined metadata provides useful supervision beyond species labels alone. This supports the broader claim that improving bioacoustic LALMs is not only a matter of adding more labeled recordings, but also of expanding the kinds of biological supervision available to the model. Captioning remains weak in absolute terms, indicating that open-ended biological description remains harder than discriminative metadata prediction.

5.3 Tier 3: Scene Understanding

Table 5 shows that ROOTS Tier 3 training mainly improves the parsing of scenes into events and relations. The largest gains occur when species identities must be bound to event-level

Table 6: BEANS-Next Tier 4 (in-context learning) results. Gibbons and DCASE use macro-F1 averaged over reference labels (A–C and A–D, respectively), with no detection represented by an empty set; other tasks use accuracy. Scores are on a 0–1 scale; higher is better. NatureLM-audio v1.0 does not accept multiple audio inputs.

	Few-shot task				
Model	Gibbons	Otters	DCASE	Crows	Unseen
AF-Next	0.007	0.255	0.119	0.225	0.230
Qwen3-Omni	0.175	0.535	0.327	0.620	0.480
NatureLM-audio v1.1	0.176	0.200	0.048	0.240	0.270
NatureLM-audio T4	**0.243**	**0.645**	**0.452**	**0.715**	**0.650**

attributes, such as which species vocalized first, which vocalization was highest or lowest in pitch, and whether vocalizations overlapped, with consistent gains across all multiple-choice characteristic tasks. Because the synthetic scenes explicitly control species identity, timing, overlap, duration, and frequency range, they provide direct supervision for this kind of structured reasoning. The supervision is nonetheless incomplete: counting remains difficult, with a species-count error slightly worse than the base models, and open-ended duration-based identification does not improve. One hypothesis is that the architecture limits fine-grained temporal perception, consistent with recent moves toward denser audio embeddings and timestamp-based positional encodings (Ghosh et al., 2026).

5.4 Tier 4: In-Context Learning

Table 6 shows the strongest evidence that ROOTS can change how the model uses its inputs. Audio Flamingo Next and the NatureLM-audio v1.1 base model remain near chance on the matching tasks, suggesting that they do not use the support audio as task-defining context. Qwen3-Omni performs well above chance on call-type and unseen-species matching, indicating that general-domain in-context learning partially transfers to bioacoustics. Tier 4 training yields the best results on all five tasks, showing that synthetic multi-audio supervision teaches the model to compare a query against in-context examples rather than rely only on labels seen in training. Gains are smallest on Gibbons, a low-SNR dataset, suggesting limited robustness to recording conditions and to the domain shift from synthetic data.

6 Conclusion

Existing audio-language models, both general-purpose and bioacoustic, cover only a limited portion of the capabilities ethologists and bioacousticians actually rely on. By organizing evaluation and training around a shared capability taxonomy, we are able to measure these gaps and show that targeted data construction—curated metadata, LLM-based text synthesis, and synthetic audio pipelines—can begin to close them. We see this as a step toward audio-language models that support the full range of bioacoustic workflows rather than only naming what is present. Substantial headroom remains, and we hope the released benchmark and dataset help the community make progress on these.

These capabilities also raise concerns worth flagging. Both BEANS-Next and ROOTS inherit taxonomic and geographic biases from public bioacoustic archives, which may limit reliability for underrepresented taxa and ecosystems. Stronger detection of endangered or vulnerable species could be misused when combined with geolocation metadata, a risk we partially mitigate by excluding endangered species from our Xeno-canto sources. And as audio-language models are deployed more widely in ecological workflows, errors—including hallucinated species, behaviors, or counts—could propagate into downstream scientific or conservation decisions. We therefore recommend these models be used as assistive tools, with predictions validated by domain experts.

AI use statement

We used generative AI tools (Gemma 4 26B MoE) as a core component of our data construction pipelines, specifically synthesizing question-answer pairs and descriptions from semi-structured and unstructured metadata, generating spectrogram-derived acoustic descriptions, and scoring candidate examples via an LLM-as-a-judge filtering step (instruction following, hallucination, audibility, and difficulty), as described in Section 3.2 and Appendix B.2. One of the authors manually reviewed a sample of pre- and post-filtered LLM outputs for each category of data as an additional quality-control step.

We also used AI coding assistants to help implement portions of the training, evaluation, data processing, and data synthesis code, and to assist with routine cluster job management and orchestration. All code was reviewed and tested by at least one of the authors before use in reported experiments.

Generative AI was used to refine some of the writing in the paper, after which it was carefully reviewed for accuracy. It was also used to spot any obvious omissions or mistakes before submission, to generate visual elements for Figure 1 and 2, and to edit the format of tables, in which case a manual review of table values was conducted. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

Reproducibility statement

We release the following artifacts to support the reproducibility of our results. BEANS-Next, ROOTS, and its synthetic audio subsets (Synthetic Strong Detection, Synthetic Detect-Diarize, and AnimalSpeak-PseudoVox) are available on Hugging Face (`https://hf.co/iclr2027anon`). For ROOTS, we do not redistribute audio. However, all audio used is publicly available, and we provide guidance for retrieving and preprocessing it accordingly. The supplementary material contains the BEANS-Next evaluation library, which implements data loading, inference servers for all evaluated models, answer parsing, and metric computation. Benchmark construction, task definitions, sample counts, metrics, and held-out taxa are described in Appendix D, including Table D.1.1 and Appendix D.2. ROOTS construction, covering templated and LLM-based tasks, filtering, and synthetic audio generation, is described in Appendix B. Templates are provided on Hugging Face with ROOTS, as well as their mapping to specific samples. Training and fine-tuning hyperparameters, augmentations, and compute are given in Appendices C.1 and C.2. The evaluation protocol, post-processing, and metric definitions are in Appendix E, and the text-only and noise controls are in Appendix E.6.

References

Paul Best, Marcelo Araya-Salas, Axel G. Ekström, Bárbara Freitas, Frants H. Jensen, Arik Kershenbaum, Adriano R. Lameira, Kenna D. S. Lehmann, Pavel Linhart, Robert C. Liu, Malavika Madhavan, Andrew Markham, Marie A. Roch, Holly Root-Gutteridge, Martin Šálek, Grace Smith-Vidaurre, Ariana Strandburg-Peshkin, Megan R. Warren, Matthew Wijers, and Ricard Marxer. Bioacoustic fundamental frequency estimation: a cross-species dataset and deep learning baseline. *Bioacoustics*, 0(0):1–28, 2025. doi: 10.1080/09524622.2025.2500380. URL `https://doi.org/10.1080/09524622.2025.2500380`.

Jack W Bradbury and Sandra Lee Vehrencamp. *Principles of animal communication*. Sinauer Associates Sunderland, MA, 1998.

Forrest Briggs, Balaji Lakshminarayanan, Lawrence Neal, Xiaoli Z. Fern, Raviv Raich, Sarah J. K. Hadley, Adam S. Hadley, and Matthew G. Betts. Acoustic classification of multiple simultaneous bird species: A multi-instance multi-label approach. *The Journal of the Acoustical Society of America*, 131(6):4640–4650, 2012. doi: 10.1121/1.4707424. URL `https://doi.org/10.1121/1.4707424`.

Alexander Brown, James Montgomery, and Saurabh Garg. Automatic construction of accurate bioacoustics workflows under time constraints using a surrogate model. *Applied*

Soft Computing, 113:107944, 2021. doi: 10.1016/j.asoc.2021.107944. URL https://doi.org/10.1016/j.asoc.2021.107944.

Jules Cauzinille, Benoît Favre, Ricard Marxer, and Arnaud Rey. Applying machine learning to primate bioacoustics: Review and perspectives. *American Journal of Primatology*, 86 (10), 2024. doi: 10.1002/ajp.23666. URL https://doi.org/10.1002/ajp.23666.

Mustafa Chasmai, Alexander Shepard, Subhransu Maji, and Grant Van Horn. The inaturalist sounds dataset. In *Advances in Neural Information Processing Systems 37*, pp. 132524–132544. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2024. doi: 10.52202/079017-4213. URL https://doi.org/10.52202/079017-4213.

Yunfei Chu et al. Qwen2-audio technical report, 2024. URL https://arxiv.org/abs/2407.10759.

Emmett F Culhane, Alice Della Penna, Roland Proud, and Camrin D Braun. Image-based machine learning reveals structural "fingerprints" of pelagic seascapes in shipboard echograms. *ICES Journal of Marine Science*, 82(12):fsaf221, 12 2025. ISSN 1054-3139. doi: 10.1093/icesjms/fsaf221. URL https://doi.org/10.1093/icesjms/fsaf221.

Maddie Cusimano, Benjamin Hoffman, Carlos Guzón-García, Angela Mcenerey, Margaret Bezrutczyk, Jen-Yu Liu, Logan S James, Sara C Keen, Louis Mahon, Olivier Pietquin, et al. Repertoire-wide contextual mapping reveals signal functions in cooperatively breeding crows. *bioRxiv*, pp. 2026–04, 2026.

Tom Denton, Scott Wisdom, and John R. Hershey. Improving bird classification with unsupervised sound separation. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 636–640. IEEE, 2022. doi: 10.1109/icassp43922.2022.9747202. URL https://doi.org/10.1109/icassp43922.2022.9747202.

Clement Farabet and Olivier Lacombe. Gemma 4: Byte for byte, the most capable open models. https://blog.google/innovation-and-ai/technology/developers-tools/gemma-4/, 2026. Accessed: 2026-04-30.

Julia Fischer, Rahel Noser, and Kurt Hammerschmidt. Bioacoustic field research: A primer to acoustic analyses and playback experiments with primates. *American Journal of Primatology*, 75(7):643–663, 2013. doi: 10.1002/ajp.22153. URL https://doi.org/10.1002/ajp.22153.

Burooj Ghani, Tom Denton, Stefan Kahl, and Holger Klinck. Global birdsong embeddings enable superior transfer learning for bioacoustic classification. *Scientific Reports*, 13(1), 2023. doi: 10.1038/s41598-023-49989-z. URL https://doi.org/10.1038/s41598-023-49989-z.

Sreyan Ghosh, Arushi Goel, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang gil Lee, Chao-Han Huck Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, and Bryan Catanzaro. Audio flamingo 3: Advancing audio intelligence with fully open large audio language models. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2025.

Sreyan Ghosh, Arushi Goel, Kaousheik Jayakumar, Lasha Koroshinadze, Nishit Anand, Zhifeng Kong, Siddharth Gururani, Sang-gil Lee, Jaehyeon Kim, Aya Aljafari, Chao-Han Huck Yang, Sungwon Kim, Ramani Duraiswami, Dinesh Manocha, Mohammad Shoeybi, Bryan Catanzaro, Ming-Yu Liu, and Wei Ping. Audio Flamingo Next: Next-Generation Open Audio-Language Models for Speech, Sound, and Music, April 2026. URL http://arxiv.org/abs/2604.10905. arXiv:2604.10905 [cs].

Sreyan Ghosh et al. Gama: A large audio-language model with advanced audio understanding and complex reasoning abilities. *arXiv preprint arXiv:2406.11768*, 2024.

Yuan Gong, Hongyin Luo, Alexander H. Liu, Leonid Karlinsky, and James Glass. Listen, think, and understand. In *International Conference on Learning Representations (ICLR)*, Vienna, Austria, 2024.

Masato Hagiwara. Aves: Animal vocalization encoder based on self-supervision. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1–5. IEEE, 2023. doi: 10.1109/icassp49357.2023.10095642. URL https://doi.org/10.1109/icassp49357.2023.10095642.

Masato Hagiwara, Benjamin Hoffman, Jen-Yu Liu, Maddie Cusimano, Felix Effenberger, and Katie Zacarian. Beans: The benchmark of animal sounds. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1–5. IEEE, 2023. doi: 10.1109/icassp49357.2023.10096686. URL https://doi.org/10.1109/icassp49357.2023.10096686.

Jenny Hamer, Eleni Triantafillou, Bart van Merriënboer, Stefan Kahl, Holger Klinck, Tom Denton, and Vincent Dumoulin. Birb: A generalization benchmark for information retrieval in bioacoustics, 2023. URL https://arxiv.org/abs/2312.07439.

Benjamin Hoffman, David Robinson, Marius Miron, Vittorio Baglione, Daniela Canestrari, Damian Elias, Eva Trapote, Felix Effenberger, Maddie Cusimano, Masato Hagiwara, et al. Synthetic data enables context-aware bioacoustic sound event detection, 2025. URL https://arxiv.org/abs/2503.00296.

Marla M Holt, Dawn P Noren, Val Veirs, Candice K Emmons, and Scott Veirs. Speaking up: Killer whales (orcinus orca) increase their call amplitude in response to vessel noise. *The Journal of the Acoustical Society of America*, 125(1):EL27–EL32, 2009.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

Mayuri Kate and Suresh Neethirajan. Big data approaches to bovine bioacoustics: a fair-compliant dataset and scalable ml framework for precision livestock welfare. *Frontiers Big Data*, 8, 2025. doi: 10.3389/fdata.2025.1723155. URL https://doi.org/10.3389/fdata.2025.1723155.

Arik Kershenbaum, Çağlar Akçay, Lakshmi Babu-Saheer, Alex Barnhill, Paul Best, Jules Cauzinille, Dena Clink, Angela Dassow, Emmanuel Dufourq, Jonathan Growcott, et al. Automatic detection for bioacoustic research: a practical guide from and for biologists and computer scientists. *Biological Reviews*, 100(2):620–646, 2025.

Jaeyeon Kim, Heeseung Yun, Tony Woo, Chao-Han Huck Yang, and Gunhee Kim. Wow-bench: Evaluating fine-grained acoustic perception in audio-language models via marine mammal vocalizations. In *Annual Meeting of the Association for Computational Linguistics*, pp. 31208–31229, 2026. doi: 10.18653/v1/2026.findings-acl.1562. URL https://doi.org/10.18653/v1/2026.findings-acl.1562.

Henry Kyalo, Henri Tonnang, James Egonyu, John Olukuru, Chrysantus Tanga, and Kennedy Senagi. Automatic synthesis of insects bioacoustics using machine learning: a systematic review. *International Journal of Tropical Insect Science*, 45(1):101–120, 2025. doi: 10.1007/s42690-024-01406-2.

Jinhua Liang, Ines Nolasco, Burooj Ghani, Huy Phan, Emmanouil Benetos, and Dan Stowell. Mind the domain gap: A systematic analysis on bioacoustic sound event detection. In *2024 32nd European Signal Processing Conference (EUSIPCO)*, pp. 1257–1261. IEEE, 2024. doi: 10.23919/eusipco63174.2024.10714948. URL https://doi.org/10.23919/eusipco63174.2024.10714948.

Zihan Liu, Zhikang Niu, Qiuyang Xiao, Zhisheng Zheng, Ruoqi Yuan, Yuhang Zang, Yuhang Cao, Xiaoyi Dong, Jianze Liang, Xie Chen, Leilei Sun, Dahua Lin, and Jiaqi Wang. Star-bench: Probing deep spatio-temporal reasoning as audio 4d intelligence. *International Conference on Learning Representations (ICLR)*, 2026.

Ziyang Ma, Yinghao Ma, Yanqiao Zhu, Chen Yang, Yi-Wen Chao, Ruiyang Xu, Wenxi Chen, Yuanzhe Chen, Zhuo Chen, Jian Cong, Kai Li, Keliang Li, Siyou Li, Xinfeng Li, Xiquan Li, Zheng Lian, Yuzhe Liang, Minghao Liu, Zhikang Niu, Tianrui Wang, Wang

Yuping, Yuxuan Wang, Yihao Wu, Guanrou Yang, Jianwei Yu, Ruibin Yuan, Zhisheng Zheng, Ziya Zhou, Haina Zhu, Wei Xue, Emmanouil Benetos, Kai Yu, Eng-Siong Chng, and Xie Chen. Mmar: A challenging benchmark for deep reasoning in speech, audio, music, and their mix. In *Advances in Neural Information Processing Systems 38*, pp. 74837–74868. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2025. doi: 10.52202/085713-2252. URL https://doi.org/10.52202/085713-2252.

Louis Mahon, Benjamin Hoffman, Logan S. James, Maddie Cusimano, Masato Hagiwara, Sarah C. Woolley, Felix Effenberger, Sara Keen, Jen-Yu liu, and Olivier Pietquin. Robust detection of overlapping bioacoustic sound events. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2025. URL https://openreview.net/forum?id=b7x4CtWynM.

Tiago A. Marques, Len Thomas, Stephen W. Martin, David K. Mellinger, Jessica A. Ward, David J. Moretti, Danielle Harris, and Peter L. Tyack. Estimating animal population density using passive acoustics. *Biological Reviews*, 88(2):287–309, 2013. doi: 10.1111/brv.12001. URL https://doi.org/10.1111/brv.12001.

Félix Michaud, Jérôme Sueur, Maxime Le Cesne, and Sylvain Haupert. Unsupervised classification to improve the quality of a bird song recording dataset. *Ecological Informatics*, 74:101952, 2023. doi: 10.1016/j.ecoinf.2022.101952. URL https://doi.org/10.1016/j.ecoinf.2022.101952.

Marius Miron, David Robinson, Milad Alizadeh, Ellen gilsenan mcmahon, Gagan Narula, Emmanuel Chemla, Maddie Cusimano, Felix Effenberger, Masato Hagiwara, Benjamin Hoffman, Sara Keen, Diane Kim, Jane K. Lawton, Jen-Yu liu, Aza Raskin, Olivier Pietquin, and Matthieu Geist. Avex: What matters for animal vocalization encoding. *International Conference on Learning Representations (ICLR)*, 2026.

Museum für Naturkunde Berlin. Animal sound archive. https://doi.org/10.15468/0bpalr, 2013. Accessed via gbif.org 2023-05-09.

Daniel A. Nieto-Mora, Susana Rodríguez-Buritica, Paula Rodríguez-Marín, J. D. Martínez-Vargaz, and Claudia Isaza-Narváez. Systematic review of machine learning methods applied to ecoacoustics and soundscape monitoring. *Heliyon*, 9(10):e20275, 2023. doi: 10.1016/j.heliyon.2023.e20275.

Ines Nolasco, Shubhr Singh, Veronica Morfi, Vincent Lostanlen, Ariana Strandburg-Peshkin, Ester Vidaña-Vila, Lisa Gill, Hanna Pamuła, Helen Whitehead, Ivan Kiskin, Frants H. Jensen, Joe Morford, Michael G. Emmerson, Elisabetta Versace, Emily Grout, Haohe Liu, Burooj Ghani, and Dan Stowell. Learning to detect an animal sound from five examples. *Ecological Informatics*, 77:102258, 2023. doi: 10.1016/j.ecoinf.2023.102258. URL https://doi.org/10.1016/j.ecoinf.2023.102258.

Karan J. Odom, Marcelo Araya-Salas, Janelle L. Morano, Russell A. Ligon, Gavin M. Leighton, Conor C. Taff, Anastasia H. Dalziell, Alexis C. Billings, Ryan R. Germain, Michael Pardo, Luciana Guimarães de Andrade, Daniela Hedwig, Sara C. Keen, Yu Shiu, Russell A. Charif, Michael S. Webster, and Aaron N. Rice. Comparative bioacoustics: a roadmap for quantifying and comparing animal sounds across diverse taxa. *Biological Reviews*, 96(4):1135–1159, 2021. doi: https://doi.org/10.1111/brv.12695. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/brv.12695.

Cristian Pérez-Granados, Jon Morant, Kevin FA Darras, Oscar H Marín-Gómez, Irene Mendoza, Miguel A Muñoz-Mohedano, Eduardo Santamaría-García, Giulia Bastianelli, Alba Márquez-Rodríguez, Michał Budka, et al. Wabad: A world annotated bird acoustic dataset for passive acoustic monitoring. *Ecology*, 107(2), 2026. doi: 10.1002/ecy.70317.

Lukas Rauch, René Heinrich, Ilyass Moummad, Alexis Joly, Bernhard Sick, and Christoph Scholz. Can masked autoencoders also listen to birds? *Transactions on Machine Learning Research*, 2025, 2025a. URL https://openreview.net/forum?id=GIBWROXo2J.

Lukas Rauch, Raphael Schwinger, Moritz Wirth, René Heinrich, Denis Huseljic, Marek Herde, Jonas Lange, Stefan Kahl, Bernhard Sick, Sven Tomforde, and Christoph Scholz. Birdset: A large-scale dataset for audio classification in avian bioacoustics. In *International Conference on Learning Representations (ICLR)*, 2025b. URL https://openreview.net/forum?id=dRXxFEY8ZE.

David Robinson, Adelaide Robinson, and Lily Akrapongpisak. Transferable models for bioacoustics with human language supervision. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1316–1320. IEEE, 2024. doi: 10.1109/icassp48485.2024.10447250. URL https://doi.org/10.1109/icassp48485.2024.10447250.

David Robinson, Marius Miron, Masato Hagiwara, and Olivier Pietquin. Naturelm-audio: an audio-language foundation model for bioacoustics. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=hJVdwBpWjt.

Christian Rutz, Michael Bronstein, Aza Raskin, Sonja C. Vernes, Katherine Zacarian, and Damián E. Blasi. Using machine learning to decode animal communication. *Science*, 381 (6654):152–155, 2023. doi: 10.1126/science.adg7314. URL https://doi.org/10.1126/science.adg7314.

S. Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. Mmau: A massive multi-task audio understanding and reasoning benchmark. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=TeVAZXr3yv.

Raphael Schwinger, Paria Vali Zadeh, Lukas Rauch, Mats Kurz, Tom Hauschild, Sam Lapp, and Sven Tomforde. Foundation models for bioacoustics - a comparative review. *Ecol. Informatics*, 96:103765, 2026. doi: 10.1016/j.ecoinf.2026.103765. URL https://doi.org/10.1016/j.ecoinf.2026.103765.

Julian C. Schäfer-Zimmermann, Vlad Demartsev, Baptiste Averly, Kiran L. Dhanjal-Adams, Mathieu Duteil, Gabriella Gall, Marius Faiß, Lily Johnson-Ulrich, Dan Stowell, Marta B. Manser, Marie A. Roch, and Ariana Strandburg-Peshkin. animal2vec and meerkat: A self-supervised transformer for rare-event raw audio input and a large-scale reference dataset for bioacoustics. *Methods in Ecology and Evolution*, 17(3):875–888, 2026. doi: 10.1111/2041-210x.70218. URL https://doi.org/10.1111/2041-210x.70218.

Dan Stowell. Computational bioacoustics with deep learning: a review and roadmap. *PeerJ*, 10:e13152, 2022. doi: 10.7717/peerj.13152. URL https://doi.org/10.7717/peerj.13152.

Jérôme Sueur, Sandrine Pavoine, Olivier Hamerlynck, and Stéphane Duvail. Rapid acoustic survey for biodiversity appraisal. *PLoS ONE*, 3(12):e4065, 2008. doi: 10.1371/journal.pone.0004065.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. Salmonn: Towards generic hearing abilities for large language models. In *International Conference on Learning Representations (ICLR)*, 2024.

Willem-Pier Vellinga and Robert Planqué. The xeno-canto collection and its relation to sound recognition and classification. *CLEF (Working Notes)*, 2015.

George Vengrovski, Miranda R. Hulsey-Vincent, Melissa A. Bemrose, and Timothy J. Gardner. Tweetybert: Automated parsing of birdsong through self-supervised machine learning. *Patterns*, 7(4):101491, 2026. doi: 10.1016/j.patter.2025.101491. URL https://doi.org/10.1016/j.patter.2025.101491.

Cheng Wang, Gelei Deng, Xianglin Yang, Han Qiu, and Tianwei Zhang. When audio and text disagree: Revealing text bias in large audio-language models. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 4878–4888, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.246. URL https://aclanthology.org/2025.emnlp-main.246/.

Ben Williams, Bart van Merriënboer, Vincent Dumoulin, Jenny Hamer, Abram B. Fleishman, Matthew McKown, Jill Munger, Aaron N. Rice, Ashlee Lillis, Clemency White, Catherine Hobbs, Tries Razak, David Curnick, Kate E. Jones, and Tom Denton. Using tropical reef, bird and unrelated sounds for superior transfer learning in marine bioacoustics. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 380(1928), 2025. doi: 10.1098/rstb.2024.0280. URL `https://doi.org/10.1098/rstb.2024.0280`.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025.

Yongyi Zang, Sean O'Brien, Taylor Berg-Kirkpatrick, Julian McAuley, and Zachary Novack. Are you really listening? boosting perceptual awareness in music-qa benchmarks. In *International Society for Music Information Retrieval (ISMIR) Conference*, 2025.

A Appendix

In this Appendix, we provide additional details on training and evaluation.

B ROOTS Construction

B.1 Templated Data Construction

Table 7: Datasets used for each training task, with a representative prompt.

Task	Example Prompt	Datasets
Species ID (common name)	What species is vocalizing in this audio recording? Common name?	Xeno-Canto, iNaturalist, Watkins, Animal Sound Archive
Species ID (scientific name)	What is the scientific name of the focal species in the audio?	Xeno-Canto, iNaturalist, Watkins, Animal Sound Archive
Species ID + context (common)	Given the context: '{{context}}', what is the common name for the focal species?	Xeno-Canto, iNaturalist, Animal Sound Archive
Species ID + context (scientific)	Given the context: '{{context}}', what is the scientific name for the focal species?	Xeno-Canto, iNaturalist, Animal Sound Archive
Full taxonomy + context	Given the context: '{{context}}', provide the full taxonomic classification.	Xeno-Canto, iNaturalist, Animal Sound Archive
Genus classification	What is the genus of the focal species in the audio?	Xeno-Canto, iNaturalist, Animal Sound Archive
Family classification	What is the family of the focal species in the audio?	Xeno-Canto, iNaturalist, Animal Sound Archive
Order classification	What is the order of the focal species in the audio?	Xeno-Canto, iNaturalist, Animal Sound Archive
Full taxonomy classification	Provide the full taxonomic classification for the species, starting with the phylum.	Xeno-Canto, iNaturalist, Watkins, Animal Sound Archive
Call type classification	What type of vocalization or call is this?	Xeno-Canto, Watkins, Animal Sound Archive
Call type (species-conditioned)	What type of call is the {{species_common}} making?	Xeno-Canto, Watkins, Animal Sound Archive
Call vs. song (binary)	Is this a call or a song?	Xeno-Canto
Alarm call presence (binary)	Is an alarm call present in this recording? Answer Yes or No.	Xeno-Canto, Animal Sound Archive
Alarm call + species (binary)	Is the {{species_common}} making an alarm call? Yes or No.	Xeno-Canto, Animal Sound Archive
Alarm / other call / song (3-way)	Classify this recording as one of: alarm, other_call, song.	Xeno-Canto, Animal Sound Archive
Call types present (multi-label)	List all call types present as a comma-separated list.	Xeno-Canto, Animal Sound Archive
Call type (fixed 5-label vocab)	Which are present? alarm call, flight call, begging call, song, call.	Xeno-Canto, Animal Sound Archive
Fine-grained call type	What are the fine-grained vocalization types in this recording?	Watkins
Fine call type (species-cond.)	What fine-grained vocalization types is the {{species_common}} producing?	Watkins
Life stage classification	What life stage is the animal in this recording?	Xeno-Canto, iNaturalist, Animal Sound Archive
Life stage (species-conditioned)	What is the life stage of the {{species_common}}?	Xeno-Canto, iNaturalist, Animal Sound Archive

Continued on next page

Task	Example Prompt	Datasets
Adult vs. juvenile (binary)	Is the focal species an adult or juvenile?	Xeno-Canto, iNaturalist
Sex prediction	Is the animal vocalizing male or female?	Xeno-Canto, Animal Sound Archive
Sex prediction (species-cond.)	What is the sex of the {{species_common}}?	Xeno-Canto, Animal Sound Archive
Species + behavior (common)	What species is vocalizing and what type of vocalization?	Xeno-Canto, Watkins, Animal Sound Archive
Species + behavior (scientific)	Give the scientific name and vocalization type.	Xeno-Canto, Animal Sound Archive
Species + fine call type	Identify the species and list the detailed vocalization types.	Watkins
Species + life stage (common)	What species is in this recording and what is its life stage?	Xeno-Canto, iNaturalist, Animal Sound Archive
Species + life stage (scientific)	Give the scientific name and life stage.	Xeno-Canto, iNaturalist, Animal Sound Archive
Species + behavior + life stage	Identify the species, call type, and life stage.	Xeno-Canto, Animal Sound Archive
Open detection (common name)	Identify any species present by their common name.	Xeno-Canto, iNaturalist, Watkins, Animal Sound Archive
Open detection (scientific name)	Identify any species present by their scientific name.	Xeno-Canto, iNaturalist, Watkins, Animal Sound Archive
Open detection (full taxonomy)	Provide the full taxonomic classification for any species heard.	Xeno-Canto, iNaturalist, Watkins, Animal Sound Archive
Species options (common, classif.)	Which of these is the focal species? Options: {{species_choices}}	Xeno-Canto, iNaturalist, Watkins, Animal Sound Archive
Species options (common, detect.)	Which of these species, if any, are present? {{species_choices}}	Xeno-Canto, iNaturalist, Watkins, Animal Sound Archive
Species options (scientific, classif.)	Which of these species (scientific name) is in the audio? {{species_choices}}	Xeno-Canto, iNaturalist, Watkins, Animal Sound Archive
Species options (scientific, detect.)	Which of these species, if any, are present? {{species_choices}}	Xeno-Canto, iNaturalist, Watkins, Animal Sound Archive
Context + options (common, classif.)	Context: {{context}}. Which species? {{species_choices}}	Xeno-Canto, iNaturalist
Context + options (common, detect.)	Context: {{context}}. Which, if any? {{species_choices}}	Xeno-Canto, iNaturalist
Context + options (scientific, classif.)	Context: {{context}}. Which species (scientific)? {{species_choices}}	Xeno-Canto, iNaturalist
Context + options (scientific, detect.)	Context: {{context}}. Which, if any (scientific)? {{species_choices}}	Xeno-Canto, iNaturalist
Genus options (classification)	Which of these genera matches the species? {{species_choices}}	Xeno-Canto, iNaturalist, Animal Sound Archive
Genus options (detection)	Which of these genera, if any, matches? {{species_choices}}	Xeno-Canto, iNaturalist, Animal Sound Archive
Family options (classification)	Which of these families matches the species? {{species_choices}}	Xeno-Canto, iNaturalist, Animal Sound Archive
Family options (detection)	Which of these families, if any, matches? {{species_choices}}	Xeno-Canto, iNaturalist, Animal Sound Archive
Taxonomic options (classification)	Which taxonomic classification matches? {{species_choices}}	Xeno-Canto, iNaturalist, Animal Sound Archive
Taxonomic options (detection)	Which classification, if any, matches? {{species_choices}}	Xeno-Canto, iNaturalist, Animal Sound Archive
Multi-label species (scientific)	List the scientific names of all species vocalizing in this clip.	Xeno-Canto
Multi-label species (common)	List the common names of all species vocalizing in this clip.	Xeno-Canto

Continued on next page

Task	Example Prompt	Datasets
Multi-label species + count	How many species are vocalizing, and what are they (scientific)?	WABAD, SSW
Multi-label species (soundscape)	List the scientific names of all species in this soundscape.	WABAD, SSW
Audio captioning (general)	Caption this audio with a rich, detailed description.	AudioSet
Bioacoustic captioning (XC synth.)	Caption the audio, using common names for any animal species.	Xeno-Canto
Bioacoustic captioning (iNat synth.)	Caption the audio, using common names for any animal species.	iNaturalist
Bioacoustic captioning (iNat short)	Give a short caption for this recording, using common names.	iNaturalist
Bioacoustic captioning (iNat sci.)	Give a short caption, using scientific names for any animal species.	iNaturalist
Call count per species	How many calls from each species? Give scientific names.	WABAD
Total call count	How many individual vocalizations can you detect in this audio?	WABAD
Temporal species ordering	List the species in the order they first vocalize (scientific).	WABAD
Frequency range prediction	What is the overall frequency range of the vocalizations?	WABAD
Per-species frequency ranges	What is the frequency range of each species' vocalizations?	WABAD
Species summary (count + freq.)	Describe each species with call counts and frequency ranges.	WABAD, DCLDE 2026
AudioSet non-bio multilabel detection	Which of these non-animal sounds are present? {{audioset_option_choices}}	AudioSet
F0 summary statistics	What is the mean fundamental frequency of this vocalization?	F0 Bioacoustic
F0 species-conditioned description	Identify the species and describe the F0 range and duration.	F0 Bioacoustic
Bird presence (binary)	Is there a bird vocalizing in this recording? Yes or No.	Xeno-Canto, Watkins
Marine mammal presence (binary)	Are there whale or dolphin sounds in this recording? Yes or No.	Watkins, Xeno-Canto, Animal Sound Archive
Mammal presence (binary)	Does this recording contain mammal vocalizations? Yes or No.	Watkins, Animal Sound Archive, iNaturalist, Xeno-Canto
Insect presence (binary)	Does this recording contain insect sounds? Yes or No.	Xeno-Canto, iNaturalist, Watkins
Amphibian presence (binary)	Does this recording contain amphibian vocalizations? Yes or No.	Xeno-Canto, iNaturalist, Watkins
Animal presence (binary)	Does this recording contain animal vocalizations? Yes or No.	Xeno-Canto

B.2 LLM-based Data Construction

B.2.1 Overview

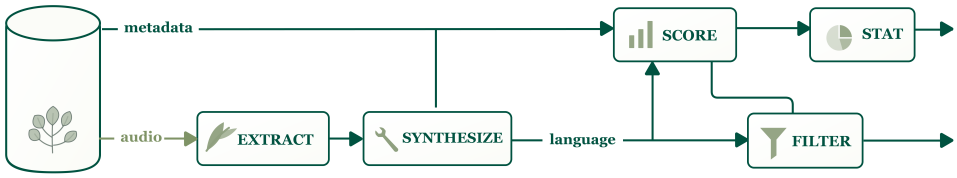

Figure 2: Overview of the LLM-based data synthesis process.

Table 8: LLM-generated datasets used for training tasks, with a representative prompt.

Task	Example Prompt	Datasets
SNR category (fixed-bin MCQ)	What is the signal-to-noise ratio (SNR) of the bird call in this recording? Options: Clean, Good, Fair, Poor.	Xeno-Canto, iNaturalist
SNR category (custom-bin MCQ)	How would you classify the SNR of the bird call in this recording? Options use LLM-defined dB ranges.	Xeno-Canto
SNR threshold (binary)	Is the SNR of the bird call in this recording above {{threshold}} dB?	Xeno-Canto
SNR prediction (open-ended)	Estimate the signal-to-noise ratio (SNR) in dB of the sound event in this recording.	Xeno-Canto
Vocalization description MCQ	Which of the following best describes the vocalization in this recording? Options: {{acoustic_descriptions}}.	AnimalSpeak Pseudovox
F0-aware vocalization description MCQ	Which description best matches the sound in this recording? Options include qualitative acoustic character and F0 information.	F0 Bioacoustic
Field-note acoustic description MCQ	Which option most accurately describes the acoustic character of this sound? Options: {{acoustic_descriptions}}.	Xeno-Canto, iNaturalist
Behavior / vocalization type MCQ (Hard)	What type of vocalization is the {{species}} producing in this recording? Options: {{behavior_choices}}.	Xeno-Canto, iNaturalist
Semantic captioning	What can you hear in this recording?	Xeno-Canto, iNaturalist
Vocalization location MCQ	At what point in the clip can you hear {{species}}? Options: near the start, middle, near the end, or absent.	WABAD, Synthetic Strong Detection, Synthetic Detect Diarize
Vocal co-occurrence (binary)	Do {{species_a}} and {{species_b}} vocalize at the same time in this recording?	WABAD, Synthetic Strong Detection, Synthetic Detect Diarize
Vocal dominance (open-ended)	Which species vocalizes the most in this recording?	WABAD, Synthetic Strong Detection, Synthetic Detect Diarize
Vocal dominance MCQ	Which of these species calls most frequently in this clip? Options: {{species_choices}}.	WABAD, Synthetic Strong Detection, Synthetic Detect Diarize
Species by vocalization order (open-ended)	Which species produces the {{ordinal}} vocalization in this recording?	WABAD, Synthetic Strong Detection, Synthetic Detect Diarize

Continued on next page

Task	Example Prompt	Datasets
Species by vocalization order MCQ	Which of the following species calls {{ordinal}} in this clip? Options: {{species_choices}}.	WABAD, Synthetic Strong Detection, Synthetic Detect Diarize
Highest-pitched species (open-ended)	Which species produces the highest-pitched vocalization in this recording?	WABAD
Highest-pitched species MCQ	Which of these animals produces the highest-pitched vocalization in this clip? Options: {{species_choices}}.	WABAD
Lowest-pitched species (open-ended)	Which species produces the lowest-pitched vocalization in this recording?	WABAD
Lowest-pitched species MCQ	Which of these animals produces the lowest-pitched vocalization in this clip? Options: {{species_choices}}.	WABAD
Longest vocalization species (open-ended)	Which species produces the longest vocalization in this recording?	WABAD, Synthetic Strong Detection, Synthetic Detect Diarize
Longest vocalization species MCQ	Which of these species makes the longest call in this clip? Options: {{species_choices}}.	WABAD, Synthetic Strong Detection, Synthetic Detect Diarize
Structural captioning	Describe the temporal structure and event organization in this recording.	WABAD, Synthetic Strong Detection, Synthetic Detect Diarize

B.3 Synthetic Audio Creation

Synthetic scenes were constructed by mixing extracted vocalizations ("pseudovox") into real environmental backgrounds. Pseudovox were obtained by source-separating Xeno-canto recordings using BirdMixIT (Denton et al., 2022) and extracting short activity windows exactly as proposed in DRASDIC (Hoffman et al., 2025); we used a filtered subset of these clips, each tagged with a species label (predicted by sl-efficientnet-bio (Miron et al., 2026) and kept where it matched the metadata), an acoustic cluster ID reused from DRASDIC, and a duration. Background recordings were drawn from SilentCities. To construct a scene, we sample a background, optionally mix two backgrounds for acoustic diversity, and insert a configurable number of pseudovox events at target SNRs sampled from a Gaussian mixture (range -20 dB to $+15$ dB), with random resampling, optional time reversal, and randomized inter-event gaps. All audio is rendered at 16 kHz. Four datasets were generated from this pipeline. **Call type MCQ** and **Binary call type detection** present the model with one or more example pseudovox clips of a target call type and ask it to identify or detect that call type in a separate audio clip. **Same/different call type** presents two pseudovox clips and asks whether they share a call type. For these three tasks, "call type" is operationalized as the intersection of species and acoustic cluster ("fine mode"); negatives are sampled from unrelated species–cluster combinations. **Few-shot detection** is a multi-label variant: the model is given example clips for 1–4 candidate sounds plus a query scene of 3–10 s and must identify which (if any) of the candidates appear in the scene; queries optionally include a short clip of the scene's background environment to help the model distinguish target events from ambient noise. All synthetic datasets are split-safe: pseudovox come from the `train_unseen` split of Xeno-canto and are not used in evaluation.

C Training

C.1 Training Hyperparameters

The stage-1 LALM was trained on only open-ended species classification data, with a global batch size of 128 for 350k steps, at a peak lr of $1e-4$ with 25,000 steps of warmup and

cosine decay to a minimum lr of $3e-5$. The stage-2 LALM was trained for 1.5 million steps and a global batch size of 128 at a peak lr of $1e-4$ with 25,000 steps of warmup and cosine decay to a minimum lr of $3e-5$. Audio was limited to 10-second clips. This completed stage-2 formed the base model for fine-tuning.

C.2 Fine-tuning Hyperparameters

The per-tier fine-tunings were instantiated from the stage 2 checkpoint for 2 epochs, using a global batch size of 128. We used an initial learning rate of $1e^{-6}$, 15,000 warmup steps to a peak learning rate of $1e^{-4}$, and cosine decay to a minimum learning rate of $1e^{-5}$. LoRA fine-tuning was applied to the query, key, value, and output projection layers, using rank 32, $\alpha = 32$, and dropout 0.1. Audio was limited to 30-second clips.

C.3 Augmentations

We apply audio augmentations during stage-2 training to improve robustness to background noise and overlapping vocalizations. With probability 0.4, we add background noise sampled from a pool of environmental and general-audio noise corpora, including DEMAND, IDMT, TUT, UrbanSound, OrCA/ship-noise datasets, WHAM, AudioSet, and SilentCities, at an SNR sampled uniformly from $[-10, 25]$ dB. Noise datasets were curated in advance, including SilentCities being processed with sl-efficient-all model (Miron et al., 2026) to filter clips with animal vocalizations. For detection-style tasks, we additionally apply a noise-masking variant with probability 0.15, replacing the target signal with background noise and training the model to output the corresponding negative label. We also use task-aware mixup with probability 0.3, mixing up to three compatible partner recordings into a primary example with Beta-distributed weights ($\alpha = 0.4$). For open-ended detection tasks, partner labels are unioned into the target answer; for option-based detection tasks, mixed-in labels are added only when they appear among the provided candidate options, otherwise they serve as distractor background. Noise augmentation is applied across all training stages, while MixUp is only applied in stage-2 onward and only for detection tasks.

C.4 Computational resources

For each fine-tune, 8 NVIDIA H100 80GB GPUs were used. For Tier 1 and Tier 4, the fine-tuning time was approximately 1 day. For Tier 2 and Tier 3, the fine-tuning time was approximately 3 days. Thus, the total train time was approximately 64 H100 days.

C.5 Corpus statistics

Figure 3 displays the audio-language and unique audio counts from each data source.

Even though we create audio-language pairs from several datasets that have been used before to train bioacoustic audio-language models, our LLM-based text construction pipeline allows us to leverage more information, including more nuanced and detailed behavioral and acoustic descriptions. Figure 4 shows the 2-gram diversity for two of the tasks that leveraged the LLM-based synthesis: acoustic and behavioral description.

D BEANS-Next Construction

D.1 Data Curation

Tier 1 benchmark processing. For expert acoustic description matching, we constructed four-way multiple-choice questions from published carrion crow vocal repertoire descriptions. Each example presents one query recording and four candidate expert descriptions from the same species, with one correct answer. Descriptions were sourced from the original study (Cusimano et al., 2026), with small modifications: we condensed the caws caw0–5, grunt0–5Q, and grunt0–5L into one group each for evaluation, to avoid asking the model to distinguish calls primarily by duration. We changed descriptions for the few calls that

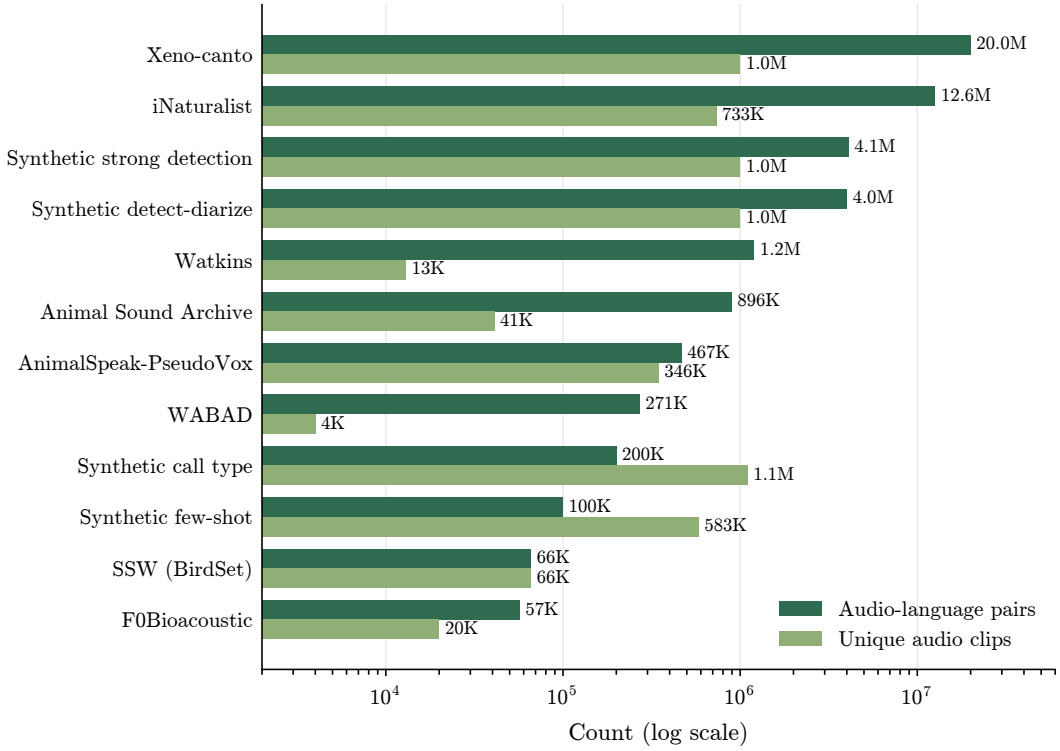

Figure 3: ROOTS data sources, with the number of audio-language pairs and unique audio clips contributed by each (log scale).

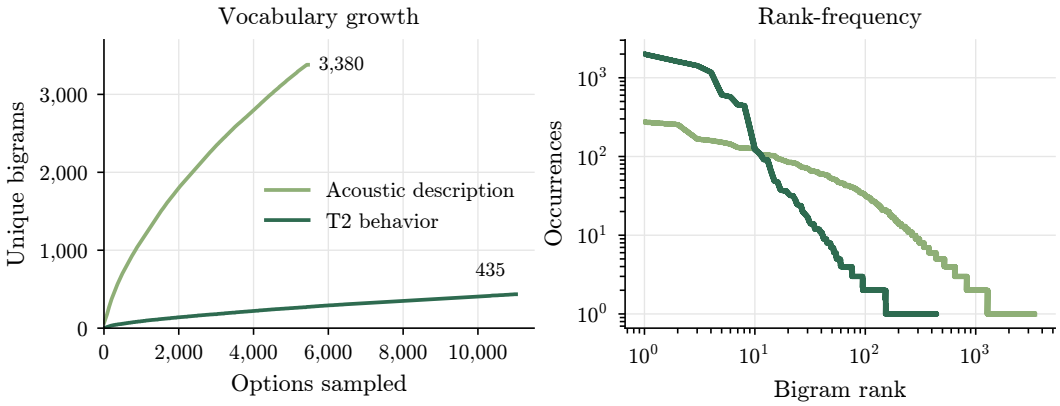

Figure 4: Lexical diversity of LLM-synthesized acoustic-description and behavior MCQ options. **Left:** number of unique word bigrams as more options are sampled. **Right:** bigram rank–frequency distribution (log–log). Acoustic descriptions keep introducing new bigrams and have a long, flat tail, whereas behavior options reuse a smaller vocabulary.

referenced another call to be standalone descriptions, and otherwise left descriptions as given by the authors.

For mean F0 prediction, we used the F0 annotation dataset by Best et al. (2025). We parsed each F0 contour, removed invalid values, computed the mean frequency over valid contour points, and rounded the target to the nearest 10 Hz. We report separate splits for taxa represented in training and a held-out split containing a single species, the spotted hyena (*Crocuta crocuta*). This split is specific to the f_0 task and separate from the species-identity held-out taxa in Appendix D.2.

For SNR prediction, we use the LLM-based pipeline to mine SNR annotations that reference the dominant vocalization in the recording from the field notes in Xeno-canto. These follow ISO 226, and Xeno-canto provides a guide for their computation, making them likely consistent.[1] We create an open-ended regression SNR evaluation, as well as an MCQ-based evaluation, that provides 4 bins with indicative SNR ranges as options.

For acoustic description, we use the LLM-based pipeline to mine acoustic descriptions from Xeno-canto sample field notes, and reformat them into captions or MCQs. The synthesis LLM retains only directly acoustic information such as pitch, texture, frequency movement, and rhythm, while excluding species names, call-type labels, behavioral context, and other semantic cues.

Tier 2 benchmark processing. For taxon presence detection, we used held-out Xeno-canto validation recordings and constructed balanced binary yes/no tasks for bird, mammal, insect, and amphibian vocalizations. Positives contain the target taxon according to focal label; negatives are sampled from other taxa in the held-out pool, with additional filtering on background species and field notes in the metadata.

For call-type evaluation, we derived binary alarm-call, flight-call, and begging-call presence tasks, plus a fixed-vocabulary multi-label task over `alarm call`, `flight call`, `begging call`, `song`, and `call`. These examples are based on BEANS-Zero call-type variants linked to Xeno-canto metadata.

We additionally include two LLM-synthesized Tier 2 semantic evaluation splits from Xeno-canto and iNaturalist BEANS-Zero/new held-out pools. Semantic captioning asks for natural-language descriptions of audible biological content, including call type, behavioral context, acoustic character, life stage when acoustically relevant, and background species when supported by the metadata. Behavior MCQs ask which audible behavior or context best matches the recording, using plausible distractors generated from the same metadata regime. These tasks differ from Tier 1 acoustic captions because semantic labels and behavioral meaning are allowed when supported by `fieldNotes`, `occurrenceRemarks`, and species metadata. We removed generations judged insufficiently audible, interesting, or difficult using an LLM-judge filter.

Tier 3 benchmark processing. For single-audio scene understanding, we used recordings from Birdeep, Powdermill, NocturnalBirdMigration, and BirdVox full-night, cropped randomly to between 1 and 30 seconds. Their ground-truth selection tables provide species labels, temporal extents, and, for Birdeep, frequency ranges; synthesis prompts convert these annotations into natural questions while excluding ambiguous rows such as ties, boundary-clipped events that could change the answer, or cases with insufficient annotation support.

Tier 3 includes LLM-based text synthesis for species counting, total vocalization counting, per-species vocalization counting, temporal referring (start, middle, or end of recording), vocalization co-occurrence, species by vocalization order (e.g. which species vocalizes third), species by highest pitch, species by lowest pitch, species by longest vocalization, species by vocalization frequency or dominance, and structural captioning, which provides an overview of vocalizations in the recording. We include both open-ended and MCQ variants where appropriate, the latter of which provide reference species as options, with a random number of them being other species that are present in the recording. We also create tasks with

[1] `https://xeno-canto.org/article/275`

templates, namely listing all species in the recording, listing the frequency range, and listing the species, frequency range, and vocalization count for each species.

Tier 4 benchmark processing. For gibbon few-shot detection, we used Hainan gibbon recordings and held out fixed support examples for three call categories: multiple-pulse calls, single-pulse calls, and duets. Each query asks which support sound, if any, appears in the target recording. We include both the full naturally imbalanced split and a balanced split for headline evaluation.

For DCASE few-shot detection, we used DCASE 2021 Task 5 sound-event examples. Each row presents audio exemplars for candidate sound classes, optionally a background clip, and a query recording. The model predicts a set of option labels or `None`. The released split is balanced between present and none examples.

For four-way call-type matching, we constructed audio-exemplar multiple-choice tasks for giant otters and carrion crows. Each example provides four labeled support clips, one per candidate call type, and a query clip from the correct class. Answer labels are balanced across A–D where possible.

For unseen-species matching, we used held-out Xeno-canto and iNaturalist taxa not included in the training mixture, based on same unseen taxa held out in NatureLM-audio (Robinson et al., 2025). Each example presents four species exemplars and a query recording; the model must select the matching species. Distractors are sampled from other held-out species, prioritizing hard, related-species negatives when available to make a more challenging task.

D.1.1 Evaluation tasks and abbreviations

Table 9: BEANS-Next evaluation task labels used in result tables. Labels include the higher-level table group followed by the abbreviated column name used in the main results.

Table label	Task	Format	Samples	Metric
Tier 1: Acoustic				
`Acoustic description / Caption`	Acoustic caption generation	Caption / OE	213	CIDEr
`Acoustic description / Desc. MCQ`	Field-note acoustic description matching	MCQ	290	Accuracy
`Acoustic description / Crow`	Carrion crow expert description matching	MCQ	200	Accuracy
`SNR / MCQ`	SNR category prediction	MCQ	607	Accuracy
`SNR / MAE (dB)`	SNR estimation	OE / regression	397	MAE (dB)
Mean f_0 / `Seen`	Mean f_0 regression for seen taxa	OE / regression	2,086	MAE (Hz)
Mean f_0 / `Held-out`	Mean f_0 regression for the held-out species	OE / regression	571	MAE (Hz)
Tier 2: Biological recognition				
`Taxon presence / Bird`	Bird presence	Binary	2,000	Accuracy
`Taxon presence / Mamm.`	Mammal presence	Binary	448	Accuracy
`Taxon presence / Amph.`	Amphibian presence	Binary	1,856	Accuracy
`Taxon presence / Insect`	Insect presence	Binary	636	Accuracy
`Call-type presence / Alarm`	Alarm-call presence	Binary	974	Accuracy
`Call-type presence / Flight`	Flight-call presence	Binary	2,000	Accuracy
`Call-type presence / Beg.`	Begging-call presence	Binary	586	Accuracy
`Call type`	Fixed-vocabulary call-type prediction	Multi-label	1,000	Macro-F1
`Behav.`	Behavior or call-type recognition	MCQ	1,169	Accuracy
`Caption`	Semantic caption generation	Caption / OE	3,000	CIDEr
Tier 3: Counting, temporal, and scene/open				

Table label	Task	Format	Samples	Metric
`Counting / #S`	Number of species present	OE / numeric	1,000	MAE
`Counting / #V`	Total number of vocalizations	OE / numeric	993	MAE
`Counting / #V/S`	Number of vocalizations per species	Open table	866	MAE
`Temporal / Referring`	Identify a vocalization from a temporal reference	MCQ	893	Accuracy
`Temporal / Overlap`	Determine whether vocalizations overlap in time	Binary	883	Accuracy
`Scene / open / Range`	Frequency range of each species in the scene	Open list	855	Species-band IoU
`Scene / open / Sum.`	Ordered summary of species activity	Caption / OE	1,000	CIDEr
`Scene / open / Caption`	Structural scene caption generation	Caption / OE	1,000	CIDEr
Tier 3: Species ID by characteristic				
`Species ID / OE / Order`	Identify species by order of vocalization	OE	942	Accuracy
`Species ID / OE / High`	Identify the highest-pitched species	OE	868	Accuracy
`Species ID / OE / Low`	Identify the lowest-pitched species	OE	803	Accuracy
`Species ID / OE / Long`	Identify the longest-vocalizing species	OE	829	Accuracy
`Species ID / OE / Freq.`	Identify the most vocally active species	OE	844	Accuracy
`Species ID / MCQ / Order`	Identify species by order of vocalization	MCQ	832	Accuracy
`Species ID / MCQ / High`	Identify the highest-pitched species	MCQ	782	Accuracy
`Species ID / MCQ / Low`	Identify the lowest-pitched species	MCQ	594	Accuracy
`Species ID / MCQ / Long`	Identify the longest-vocalizing species	MCQ	863	Accuracy
`Species ID / MCQ / Freq.`	Identify the most vocally active species	MCQ	873	Accuracy
Tier 4: Multi-audio few-shot				
`Few-shot task / Gibbons`	Gibbon few-shot detection	Few-shot multi-label	200	Macro-F1
`Few-shot task / Otters`	Giant otter call-type recognition	Multi-audio MCQ	200	Accuracy
`Few-shot task / DCASE`	DCASE few-shot detection	Few-shot multi-label	200	Macro-F1
`Few-shot task / Crows`	Carrion crow call-type recognition	Multi-audio MCQ	200	Accuracy
`Few-shot task / Unseen`	Unseen species identification	Multi-audio MCQ	200	Accuracy
Additional evals				
`CBI`	BEANS-Zero Cornell Bird Identification	OE classification	3,620	Accuracy
`HBDB`	BEANS-Zero HumBugDB mosquito classification	OE classification	1,859	Accuracy
`ESC50`	BEANS-Zero ESC-50 environmental sound classification	OE classification	400	Accuracy
`Call-type`	BEANS-Zero bird call-vs-song classification	OE classification	1,000	Accuracy
`Lifestage`	BEANS-Zero bird life-stage classification	OE classification	466	Accuracy

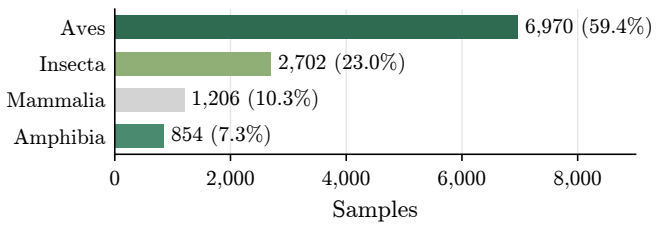

Figure 5: Taxonomic class distribution of BEANS-Next samples with full taxonomic metadata (Tiers 2 and 4).

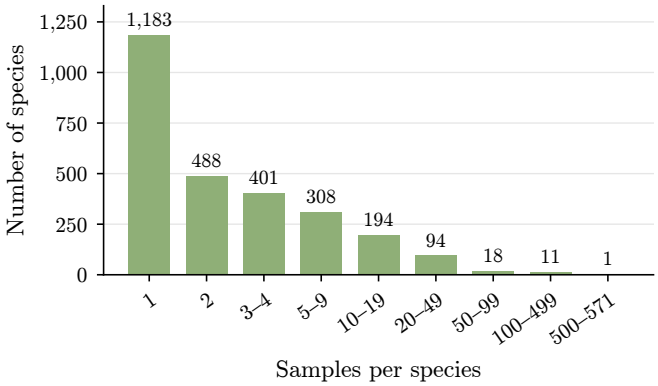

Figure 6: Species frequency distribution across BEANS-Next: number of species by samples per species. Of the 2,698 unique species, 1,183 (44%) appear only once, while a few focal species (e.g., *Crocuta crocuta*, *Corvus corone*) contribute hundreds of samples, stressing few-shot and zero-shot generalization rather than memorization of frequent taxa.

Table label	Task	Format	Samples	Metric
`N-indv`	BEANS-Zero zebra-finch individual-count classification	OE classification	1,160	Accuracy
`BirdSet`	BirdSet classification	OE classification	117,092	Accuracy

D.2 List of BEANS-Next held-out taxa

Held-out species:

1. *Aethopyga shelleyi*
2. *Arachnothera dilutior*
3. *Sitta castanea*
4. *Carpodacus rodopeplus*
5. *Aethopyga ignicauda*
6. *Pachycephala soror*
7. *Herpsilochmus roraimae*
8. *Amazona dufresniana*
9. *Metallura aeneocauda*
10. *Thlypopsis fulviceps*
11. *Monarcha frater*
12. *Kleinothraupis reyi*
13. *Aplonis magna*
14. *Phylloscopus misoriensis*
15. *Agapornis pullarius*
16. *Amazona versicolor*
17. *Saltator cinctus*
18. *Xiphocolaptes falcirostris*
19. *Passer insularis*
20. *Chalcomitra balfouri*
21. *Arremonops tocuyensis*
22. *Atlapetes meridae*
23. *Colluricincla obscura*
24. *Saltator maxillosus*
25. *Philemon meyeri*
26. *Thamnophilus insignis*
27. *Aulacorhynchus whitelianus*
28. *Sirystes subcanescens*
29. *Sporophila nigrorufa*
30. *Zoothera mollissima*
31. *Thlypopsis inornata*
32. *Picumnus spilogaster*
33. *Columba arquatrix*
34. *Petrochelidon rufocollaris*
35. *Pyrrhura griseipectus*
36. *Myiothlypis chrysogaster*
37. *Thripophaga amacurensis*
38. *Herpsilochmus motacilloides*
39. *Progne dominicensis*
40. *Heliodoxa branickii*
41. *Asthenes arequipae*
42. *Gerygone fusca*
43. *Otus thilohoffmanni*
44. *Inezia subflava*
45. *Charadrius montanus*
46. *Petroica polymorpha*
47. *Symposiachrus vidua*
48. *Dicrurus lophorinus*

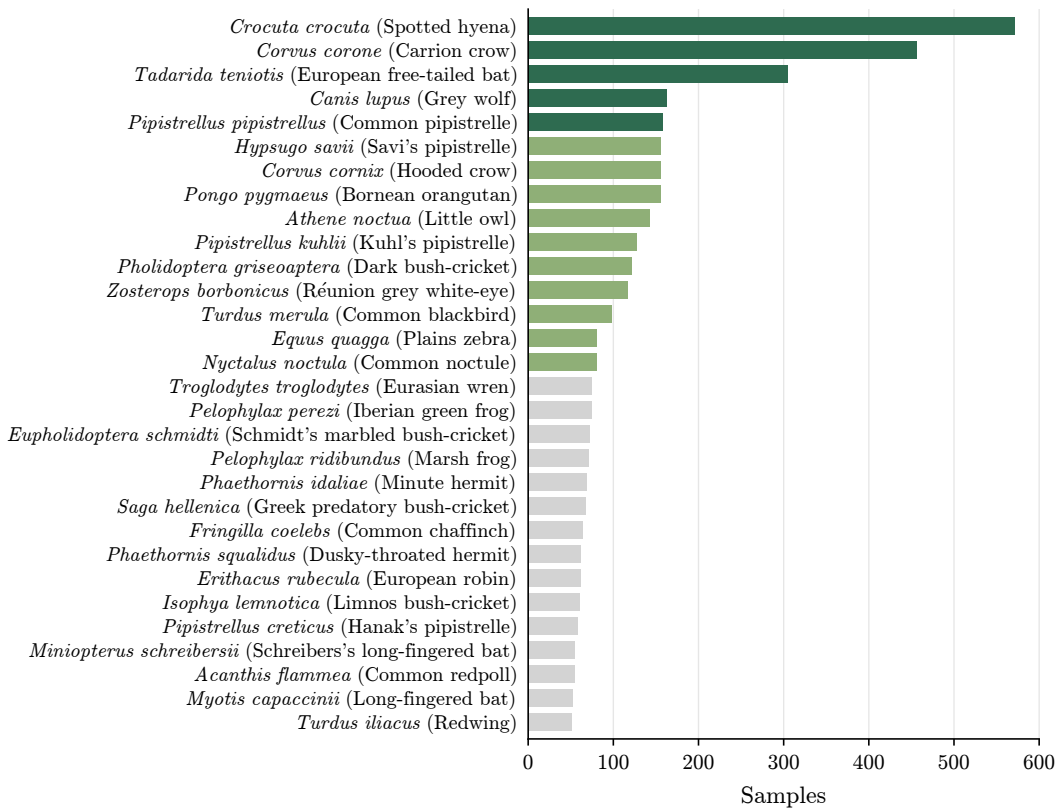

Figure 7: Top 30 species by sample count in BEANS-Next. High-count species are predominantly focal study subjects from controlled recording campaigns (e.g., *Corvus corone* from the crow call-type dataset, *Tadarida teniotis* and *Pipistrellus* spp. from bat monitoring). The rapid drop-off after the top few species illustrates the long-tail character of the overall species distribution.

Capability	BEANS	BirdSet	WoW-Bench	BEANS-Next
Species recognition	**Yes**	**Yes**	No	**Yes**
Acoustic measurement	No	No	**Yes**	**Yes**
Temporal reasoning	No	No	No	**Yes**
Relational reasoning	No	No	No	**Yes**
Multi-audio reasoning	No	No	No	**Yes**
In-context learning	No	No	No	**Yes**

Table 10: Comparison of benchmark coverage across bioacoustic audio-language capabilities. Existing benchmarks primarily evaluate species recognition or isolated acoustic perception, while BEANS-Next evaluates a broader capability space required by ethological workflows.

49. *Pycnonotus penicillatus*
50. *Melanerpes herminieri*
51. *Zosterops mysorensis*
52. *Oenanthe xanthoprymna*
53. *Artamus monachus*
54. *Caprimulgus pulchellus*
55. *Psarocolius cassini*
56. *Symposiachrus infelix*
57. *Zosterops cinereus*
58. *Circus cinereus*
59. *Geotrygon chrysia*
60. *Microspingus trifasciatus*
61. *Pternistis harwoodi*
62. *Ceblepyris caesius*
63. *Ficedula disposita*
64. *Treron affinis*
65. *Geokichla wardii*
66. *Campethera bennettii*
67. *Alcedo semitorquata*
68. *Buteo japonicus*
69. *Apus bradfieldi*
70. *Pterocles personatus*
71. *Melaniparus fringillinus*
72. *Poecile hypermelaenus*
73. *Circus buffoni*
74. *Pycnonotus blanfordi*
75. *Machlolophus aplonotus*
76. *Estrilda ochrogaster*
77. *Touit batavicus*
78. *Mirafra gilletti*
79. *Pternistis icterorhynchus*
80. *Accipiter collaris*
81. *Knipolegus lophotes*

82. *Nothoprocta taczanowskii*
83. *Pachycephala modesta*
84. *Vanellus tricolor*
85. *Caprimulgus andamanicus*
86. *Ardenna grisea*
87. *Mixornis kelleyi*
88. *Cinnyris johannae*
89. *Recurvirostra novaehollandiae*
90. *Sitta leucopsis*
91. *Petroica pusilla*
92. *Amazilia luciae*
93. *Melaniparus fasciiventer*
94. *Egretta picata*
95. *Columba pollenii*
96. *Rallus madagascariensis*
97. *Heliodoxa gularis*
98. *Carpodacus roseus*
99. *Zosterops chloronothos*
100. *Pachycephala lorentzi*
101. *Saucerottia cyanura*
102. *Cinclosoma marginatum*
103. *Bucco noanamae*
104. *Certhia nipalensis*
105. *Pachycephala lanioides*
106. *Carpodacus trifasciatus*
107. *Chorthippus acroleucus*
108. *Chlidonias albostriatus*
109. *Hirundo domicola*
110. *Falco concolor*
111. *Dryocopus schulzii*
112. *Rhyticeros undulatus*
113. *Quiscalus nicaraguensis*
114. *Cisticola brunnescens*
115. *Knipolegus cyanirostris*
116. *Ardenna carneipes*
117. *Lybius rubrifacies*
118. *Climacteris melanurus*
119. *Puffinus opisthomelas*
120. *Manorina melanotis*
121. *Celebesica abbotti*
122. *Otus mayottensis*
123. *Trachyphonus margaritatus*
124. *Oenanthe dubia*
125. *Chloropsis flavipennis*
126. *Ploceus alienus*
127. *Phalacrocorax varius*
128. *Ploceus pelzelni*
129. *Merops mentalis*
130. *Passer gongonensis*
131. *Myzomela cineracea*
132. *Pachycephala feminina*
133. *Brachypteryx sinensis*
134. *Lonchura flaviprymna*
135. *Ninox natalis*
136. *Myrmelastes caurensis*
137. *Buteo trizonatus*
138. *Apalis chariessa*
139. *Ficedula nigrorufa*
140. *Pica mauritanica*
141. *Anthreptes reichenowi*
142. *Sholicola major*
143. *Vireo osburni*
144. *Anas capensis*
145. *Ducula luctuosa*
146. *Lanius newtoni*
147. *Odontophorus dialeucos*
148. *Bostrychia olivacea*
149. *Cinnyris tsavoensis*
150. *Ploceus heuglini*
151. *Myzomela nigrita*
152. *Falco cherrug*
153. *Ixobrychus sturmii*
154. *Rhipidura semirubra*
155. *Haematopus chathamensis*
156. *Anthus brachyurus*
157. *Oenanthe lugens*
158. *Columba rupestris*
159. *Rhyticeros subruficollis*
160. *Zosterops vellalavella*
161. *Anthus sokokensis*
162. *Phaethornis idaliae*
163. *Picus dedemi*
164. *Muscicapa segregata*
165. *Cyanomitra bannermani*
166. *Polioptila facilis*
167. *Platysteira albifrons*
168. *Dicaeum pygmaeum*
169. *Puffinus assimilis*
170. *Rhipidura kubaryi*
171. *Ploceus katangae*
172. *Canis lupaster*
173. *Hyla andersonii*
174. *Ranoidea nudidigita*
175. *Ranoidea aurea*
176. *Litoria tyleri*
177. *Dendropsophus joannae*
178. *Okanagana occidentalis*
179. *Litoria latopalmata*
180. *Magicicada tredecassini*
181. *Orchelimum silvaticum*
182. *Oecanthus celerinictus*
183. *Empidonomus aurantioatrocristatus*
184. *Bufotes boulengeri*
185. *Oecanthus nigricornis*
186. *Myrmothera fulviventris*
187. *Psaltoda adonis*
188. *Rana dalmatina*
189. *Dendropsophus sanborni*
190. *Hyperolius stictus*
191. *Hyperolius pictus*
192. *Hyla eximia*
193. *Leptodactylus natalensis*
194. *Oecanthus californicus*
195. *Hyperolius parallelus*
196. *Gryllus cohni*
197. *Physeter macrocephalus*
198. *Eleutherodactylus unicolor*
199. *Gryllus bermudensis*
200. *Anas penelope*

Held-out genera:

1. *Aglaeactis*
2. *Drepanorhynchus*
3. *Lesbia*
4. *Nemobius*
5. *Meconema*
6. *Pseudochorthippus*
7. *Caliechthrus*
8. *Pachycare*
9. *Rhodothraupis*
10. *Astrapia*
11. *Probosciger*
12. *Amazonetta*
13. *Ocyalus*
14. *Nandayus*
15. *Rhinocrypta*
16. *Heterocercus*
17. *Jacamaralcyon*
18. *Hymenops*
19. *Doliornis*
20. *Eugerygone*
21. *Cryptosylvicola*
22. *Taeniopygia*
23. *Catharopeza*
24. *Eurostopodus*
25. *Tylas*
26. *Vini*
27. *Ptychoramphus*
28. *Speculanas*
29. *Aphelocephala*
30. *Stipiturus*
31. *Procarduelis*
32. *Rhopophilus*
33. *Neopsephotus*
34. *Enodes*
35. *Leucocarbo*
36. *Gymnophaps*
37. *Goldmania*
38. *Oreomystis*
39. *Rhodostethia*
40. *Falcipennis*
41. *Pachycoccyx*
42. *Cryptotympana*
43. *Tympanistalna*
44. *Cyrtoxipha*
45. *Afrixalus*
46. *Uperoleia*
47. *Urocitellus*
48. *Chalcorana*
49. *Aiolopus*
50. *Speothos*

Held-out families:

1. *Elachuridae*
2. *Calyptophilidae*
3. *Pelecanoididae*
4. *Phocoenidae*
5. *Alytidae*
6. *Castoridae*
7. *Dicroglossidae*
8. *Suidae*
9. *Prophalangopsidae*
10. *Octodontidae*

E Evaluation Pipeline and Metrics

E.1 Models

NatureLM-audio v1.0. NatureLM-audio v1.0 (Robinson et al., 2025) is an audio-language foundation model specifically designed for bioacoustics, combining a BEATs audio encoder with a Llama 3.1 8b instruction-tuned language model backbone. The model was trained on paired audio-text data spanning animal vocalizations, speech, and music, and supports a broad range of zero-shot bioacoustic tasks, including species classification, vocalization detection, call-type recognition, and audio captioning. NatureLM-audio was introduced alongside the BEANS-Zero benchmark and demonstrated strong transfer performance on unseen taxa and tasks.

NatureLM-audio v1.1. We train NatureLM-audio v1.1 for the first time in this work. It retains the architecture of NatureLM-audio v1.0 but incorporates infrastructure improvements supporting longer audio and multiple audio files as input, together with newly ingested training data from the same underlying bioacoustic archives. NatureLM-audio v1.1 serves as the common base checkpoint for the tier-specific ROOTS experiments.

Audio Flamingo Next (AF-Next). Audio Flamingo Next (Ghosh et al., 2026) is a large open audio-language model for unified understanding of speech, environmental sounds, and music. AF-Next extends the Audio Flamingo family with substantially larger-scale training data, long-context audio processing (up to 30 minutes), and temporal chain-of-thought reasoning mechanisms grounded in timestamps. The model achieves state-of-the-art or competitive performance across a wide range of audio understanding and reasoning benchmarks. AF-Next builds upon earlier Audio Flamingo models that introduced few-shot and conversational audio-language reasoning capabilities for open-ended audio understanding tasks. Model checkpoints and project resources are available through NVIDIA research releases and Hugging Face.[2]

Qwen3-Omni Instruct. Qwen3-Omni (Xu et al., 2025) is a unified multimodal foundation model supporting text, image, audio, and video understanding and generation within a single architecture. The model employs a "Thinker–Talker" mixture-of-experts design intended to jointly support multimodal reasoning and streaming speech generation. The audio subsystem achieves strong performance across audio and audio-visual benchmarks, including competitive results against large proprietary systems. In our experiments, we use the instruction-tuned Qwen3-Omni variant for zero-shot audio understanding and captioning tasks. Public model implementations and checkpoints are available through the Qwen repositories.[3]

E.2 Evaluation library

Our evaluation is fully decoupled from model serving: models run behind an HTTP *predict* endpoint, while the BEANS-Next runner performs post-processing and metric computation locally. For each dataset example, the runner sends the audio (and task prompt) to the model server, receives a raw text prediction, applies a task-specific post-processing pipeline, and then computes deterministic metrics (or routes to an LLM judge for non-deterministic tasks). The evaluation writes both per-example artifacts (raw, processed, and scored predictions) and an aggregate summary with per-task metrics.

GPT-4o audio-preview and Gemini 3.8 Flash were excluded due to high rates of API-level failures and cost, respectively. GPT-4o refused to process approximately 30% of examples

[2] `https://huggingface.co/nvidia/audio-flamingo-next-hf`
[3] `https://github.com/QwenLM/Qwen3-Omni`

citing playback restrictions. For Gemini 3.8 Flash, we sampled 500 samples from all tasks in BEANS-Zero and BEANS-Next, and estimated the total cost for evaluating on the two benchmarks with 'high' thinking mode to be $7,500.

E.3 Post-processing: parsing and label matching

Let $\hat{y}_{\text{raw}}$ be the raw model text output for an example. Post-processing produces a *segment list* $S = [s_1, \ldots, s_k]$ and a canonical joined string $\hat{y}$ obtained by concatenating segments with "`, `". The pipeline is explicitly split into two phases:

Parser phase (structural extraction). Parser steps only reshape text into label-like fragments without normalization. For classification and detection tasks, we apply comma-based parsing:

- `parse_labels_comma`: splits each segment on commas into multiple segments, while avoiding splitting numeric thousands separators (e.g., "`1,654.75`").

Cleaner phase (normalization and label alignment). Cleaner steps normalize segments and align them to the task label vocabulary:

- `normalize_whitespace`: collapses runs of whitespace and strips leading/trailing spaces per segment.
- `strip_eos`: truncates any text following a model-specific end-of-sequence marker (default "`<|end_of_text|>`") per segment.
- `fuzzy_match_to_labels`: maps each segment to the closest vocabulary label by minimum Levenshtein distance; optionally, a distance threshold can reject low-confidence matches (replacing them with a default label).
- `conservative_match_to_labels`: a higher-precision variant intended for multi-label detection, using (i) case-insensitive exact match, (ii) case-insensitive longest substring match, then (iii) bounded Levenshtein matching; unmatched fragments are typically dropped.
- `extract_label_from_text` / `extract_mcq_choice_from_text` / `extract_scientific_name_from_text` / `extract_hz_bucket_from_text`: specialized extractors used by tasks whose model outputs are verbose (e.g., multiple-choice explanations, scientific-name outputs with timestamps, or numeric frequency estimates).

Fuzzy label matching is applied only when a task has a fixed label vocabulary (e.g., classification/detection). For captioning and other free-form tasks (open-ended, counting, QA), post-processing is limited to whitespace/EOS cleanup to avoid corrupting natural-language outputs.

E.4 Task types and metrics

Each evaluation task declares a task type, which determines both the post-processing configuration and the metric family.

E.4.1 Classification

Classification tasks produce a single label from a finite vocabulary. After post-processing and label alignment, we compute:

- **Accuracy**: fraction of examples where the (post-processed) predicted label exactly matches the ground-truth label.
- **Macro precision**, **macro recall**, **macro F1**: computed by averaging per-class precision/recall/F1 equally across classes.

E.4.2 Detection (multi-label presence)

Detection tasks allow multiple labels per example (presence/absence). After comma parsing and label matching, we compute:

- **Macro Average Precision (mAP)**: average precision computed per label from the precision–recall curve and then macro-averaged across labels.
- **Macro precision**, **macro recall**, **macro F1**: computed from predicted vs. target label sets (per example), then aggregated per task.

E.4.3 Captioning

Captioning tasks produce free-form text descriptions. Because single-sample CIDEr is degenerate, we compute captioning metrics at the corpus level:

- **CIDEr (corpus mean, normalized to** $[0, 1]$**)**: reported as the mean CIDEr score over the evaluation set.

E.5 Computational resources

For NatureLM-audio v1.0 and v1.1 checkpoints, as well as Audio Flamingo Next, an NVIDIA A100 40GB GPU was used for each for approximately 1 day to run all evaluations. For Qwen3-Omni, an NVIDIA H100 80GB GPU was used for approximately 1 day. These estimates are similar for the 3 audio-language models and 6 configurations of language models for the ablations in Sec. E.6. However, only BEANS-Next evaluation was done for these ablations, which accounts for around an eighth of samples. Thus, the total evaluation resources were around 4 A100 days and 1 H100 days for the main evaluations, and 1 A100 day and 0.5 H100 day for the ablations.

E.6 Text-only and Gaussian-noise controls

MMAU and music question answering have indicated that LALMs can sometimes correctly respond to queries even when the correct audio is not provided (Sakshi et al., 2025; Zang et al., 2025). Complementary experiments find that audio-language models can favor textual cues when text and audio conflict (Wang et al., 2025). We thus evaluate the base LLMs of our chosen LALMs, namely Llama 3.1 8B, Qwen2.5 7B, and Qwen3 30B-A3B (although, Qwen3-Omni utilizes a modified version of it) without audio. We also measure performance of the chosen LALMs, namely NatureLM-audio v1.0, Audio Flamingo Next (AF-Next), and Qwen3-Omni, when Gaussian noise is provided as input, rather than the correct audio file. We use these experiments as proxies to assess whether textual cues are benefiting models during evaluation in BEANS-Next.

Conditions. We distinguish two text-only conditions. In the *informed* condition, all audio inputs and placeholders are removed and the following instruction is prepended to each question:

> Audio is intentionally unavailable. Answer using only the text and answer choices. Make your best guess and follow the requested answer format. Do not refuse or explain.

In the *silent* condition, audio is removed without this instruction. The question text and answer choices are otherwise unchanged. We use each model's native chat template and deterministic decoding. For the three audio-language models, we compare real audio with duration-matched Gaussian noise at -20 dBFS RMS. Noise is supplied with the ordinary audio prompt, without an additional informed condition.

Scoring and answer validity. We apply the same task-specific scoring rules across conditions. Unambiguous answer prefixes and bulleted call-type lists are accepted; explicit missing-audio refusals are invalid answers rather than negative presence predictions. Invalid

MCQ and binary-presence answers receive zero accuracy. Regression MAE is conditional on a parseable numerical answer, with its parse rate (the share of answers with a parseable number) reported alongside the error, as in Tables 3 and 5. For per-species counting, validity requires a parseable species–count mapping or explicit absence. Table 14(b) separately reports parse rates for MCQ and label-set tasks, pooling accepted answers over attempted examples.

Results. Tables 11–14 report the control results. Informing text models that audio is unavailable substantially changes their response behavior. Across Tier 1 MCQs, the three LLMs produce valid answers on 99.5–99.6% of examples when informed, compared with 1.3–5.8% in the silent condition. For Tier 3 MCQs, the corresponding ranges are 91.6–97.5% and 0.3–1.6%. Thus, an informed–silent performance difference reflects both answer validity and the usefulness of textual priors. Conversely, a valid answer can be uninformative: silent Qwen3 and Llama answer "Yes" on all 2,000 bird-presence examples, obtaining 50% accuracy. Informed Qwen3 predicts all five call types on every example, yielding 46.4% macro-F1 but zero exact-set accuracy. This baseline explains why a substantial multi-label score need not indicate sensitivity to the recording.

Real audio nevertheless provides a clear advantage on several tasks. Qwen3-Omni bird and amphibian presence accuracy falls from 90.0% and 86.5% with real audio to 50.0% for both with noise; Audio Flamingo Next shows similar reductions, from 88.3% and 80.4% to 50.0% and 49.5%. Other comparisons are less decisive. Conditional regression errors can conceal low answer coverage: Audio Flamingo Next provides numerical estimates for only 1.8% of held-out f_0 examples with real audio, while several no-audio conditions provide none. Tier 3 results also depend on output compatibility: only 0.05% of NatureLM-audio v1.0 real-audio MCQ responses are parseable under the single-answer protocol. Its near-zero scores should therefore not be interpreted as a direct ranking of scene-understanding ability. Together, these controls show why task performance, answer validity, and simple constant-answer baselines should be considered jointly when assessing reliance on audio.

Table 11: BEANS-Next acoustic and biological category recognition controls. Rows are grouped by input: real audio, duration-matched Gaussian noise, or no audio with (*informed*) or without (*silent*) the missing-audio instruction. Metrics and formatting follow Tables 3 and 4: higher is better unless marked ↓, small grey percentages give the parse rate, MAEs based on fewer than half of the answers are shown in grey, and – marks no parseable answer. Insect and begging-call presence use 636 and 586 examples, respectively. †Retained inference failures: three Llama 3.1 8B seen-f_0 examples and one AF-Next caption per audio condition.

(a) Tier 1: acoustic perception

	Acoustic description				SNR		Mean f_0 MAE (Hz) ↓	
Model	Caption	Desc.	MCQ	Crow	MCQ	MAE (dB) ↓	Seen	Held-out
Real audio								
AF-Next	0.4		50.1	47.5	33.9	34.6 9%	5652 13%	249 2%
Qwen3-Omni	0.0		47.0	45.5	7.7	20.0 83%	3728 >99%	203 99%
NatureLM-audio v1.0	0.3		23.6	33.5	31.9	24.9 97%	4087 >99%	208 100%
Gaussian noise								
AF-Next	0.8		25.6	36.5	30.4	38.9 17%	4860 5%	–
Qwen3-Omni	0.0		13.5	25.0	4.3	42.4 52%	5440 77%	254 22%
NatureLM-audio v1.0	0.5		21.8	25.5	28.5	36.1 98%	3773 100%	418 100%
No audio, informed								
AF-Next	0.0		43.2	38.0	32.5	24.3 95%	4492 100%	208 100%
Qwen2.5 7B	0.4		29.6	32.5	28.0	37.1 91%	4235 100%	79 100%
Qwen3 30B-A3B	0.5		22.9	32.5	22.4	16.6 97%	–	–
Llama 3.1 8B	0.1		27.5	34.5	38.1	49.5 19%	–†	–
No audio, silent								
AF-Next	0.0		40.2	37.0	32.2	27.9 37%	–	–
Qwen2.5 7B	0.0		0.1	11.5	0.0	–	–	–
Qwen3 30B-A3B	0.0		0.0	30.5	0.0	19.9 <1%	–	–
Llama 3.1 8B	0.0		0.0	8.5	0.0	49.0 <1%	–	–

(b) Tier 2: biological category recognition

	Taxon presence				Call-type presence					
Model	Bird	Mamm.	Amph.	Insect	Alarm	Flight	Beg.	Call type	Behav.	Caption
Real audio										
AF-Next	88.3	49.5	80.4	75.8	59.6	55.6	54.4	44.1	44.6	0.0†
Qwen3-Omni	90.0	49.4	86.5	75.6	62.7	60.0	66.7	23.4	53.3	0.1
NatureLM-audio v1.0	63.0	58.2	64.9	55.2	51.2	54.3	50.2	49.9	58.5	0.5
Gaussian noise										
AF-Next	50.0	50.4	49.5	50.0	50.0	50.0	50.0	20.4	43.2	0.1†
Qwen3-Omni	50.0	50.0	50.0	50.2	50.0	50.0	50.0	13.7	48.7	0.1
NatureLM-audio v1.0	51.1	50.0	50.9	50.2	50.0	50.1	50.0	15.3	46.2	0.0
No audio, informed										
AF-Next	50.0	50.0	50.0	50.0	50.0	50.0	50.0	46.4	34.0	0.0
Qwen2.5 7B	50.0	50.0	50.0	50.0	50.0	50.0	50.0	13.7	31.5	0.0
Qwen3 30B-A3B	50.0	50.0	50.0	50.0	50.0	50.0	50.0	46.4	33.2	0.0
Llama 3.1 8B	50.0	50.0	50.0	50.0	50.0	50.0	50.0	28.1	31.4	0.0
No audio, silent										
AF-Next	50.0	50.0	50.0	50.0	50.0	50.0	50.0	46.4	39.2	0.1
Qwen2.5 7B	0.0	0.0	0.0	50.0	0.0	0.0	0.0	0.0	0.1	0.0
Qwen3 30B-A3B	50.0	50.0	50.0	50.0	50.0	50.0	50.0	0.0	3.2	0.0
Llama 3.1 8B	50.0	50.0	0.0	50.0	0.0	0.0	50.0	0.0	0.3	0.0

Table 12: BEANS-Next scene-understanding controls; inputs and formatting as in Table 11, metrics as in Table 5. `#V` MCQ (accuracy) and Presence (vocalization presence, accuracy) are additional Tier 3 tasks not shown in Table 5. For Range, unmatched or unparseable examples score zero. NatureLM-audio v1.0 segment-wise responses are scored as returned.

(a) Counting

	Counting MAE ↓			
Model	`#S`	`#V`	`#V/S`	`#V` MCQ
Real audio				
AF-Next	1.65 97%	4.23 >99%	3.08 53%	26.0
Qwen3-Omni	1.20 94%	3.82 >99%	1.92 85%	51.8
NatureLM-audio v1.0	1.00 41%	6.58 25%	2.11 2%	0.0
Gaussian noise				
AF-Next	2.38 93%	4.77 100%	6.08 62%	47.8
Qwen3-Omni	2.73 82%	5.24 >99%	1.29 11%	54.2
NatureLM-audio v1.0	1.11 66%	3.18 83%	–	0.1
No audio, informed				
AF-Next	1.85 100%	7.59 100%	3.31 34%	15.4
Qwen2.5 7B	2.21 100%	4.89 100%	5.37 34%	33.5
Qwen3 30B-A3B	2.03 100%	4.68 100%	2.69 3%	29.0
Llama 3.1 8B	1.94 43%	1.99 28%	2.17 33%	20.4
No audio, silent				
AF-Next	1.96 100%	5.40 100%	–	13.6
Qwen2.5 7B	2.07 30%	5.99 39%	1.94 32%	0.6
Qwen3 30B-A3B	–	–	2.04 34%	0.0
Llama 3.1 8B	1.95 51%	4.06 56%	2.37 49%	0.0

(b) Temporal and scene-level tasks

		Temporal		Scene / open		
Model	Presence	Referring	Overlap	Range	Sum.	Caption
Real audio						
AF-Next	50.9	33.5	47.3	0.0	0.0	0.0
Qwen3-Omni	51.2	25.0	48.9	0.0	0.0	0.0
NatureLM-audio v1.0	49.7	0.0	58.4	0.0	0.1	0.1
Gaussian noise						
AF-Next	50.0	18.9	55.2	0.0	0.0	0.0
Qwen3-Omni	50.0	21.4	54.9	0.0	0.0	0.1
NatureLM-audio v1.0	61.5	0.2	54.9	0.0	0.0	0.0
No audio, informed						
AF-Next	50.0	22.5	55.4	0.0	0.0	0.0
Qwen2.5 7B	50.0	21.1	54.6	0.0	0.0	0.1
Qwen3 30B-A3B	50.0	25.6	51.3	0.0	0.1	0.0
Llama 3.1 8B	50.0	22.5	54.8	0.0	0.0	0.1
No audio, silent						
AF-Next	50.0	24.9	54.4	0.0	0.0	0.0
Qwen2.5 7B	0.0	0.0	36.4	0.0	0.0	0.0
Qwen3 30B-A3B	31.0	0.3	14.7	0.0	0.0	0.0
Llama 3.1 8B	8.1	0.3	37.6	0.0	0.0	0.0

Table 13: BEANS-Next species-ID controls; inputs, metrics, and formatting as in Tables 5 and 12. †Qwen2.5 7B silent lowest-pitch identification uses 799 of 803 responses.

Species ID by characteristic

	Open-ended					MCQ				
Model	Order	High	Low	Long	Freq.	Order	High	Low	Long	Freq.
Real audio										
AF-Next	0.0	0.0	0.0	0.0	0.0	27.0	32.2	46.0	32.6	24.5
Qwen3-Omni	0.0	0.0	0.0	0.0	0.0	25.7	33.6	19.7	19.2	13.4
NatureLM-audio v1.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0
Gaussian noise										
AF-Next	0.0	0.0	0.0	0.0	0.0	26.8	31.7	40.1	30.2	23.0
Qwen3-Omni	0.0	0.0	0.0	0.0	0.0	27.6	24.0	18.2	14.0	10.5
NatureLM-audio v1.0	0.1	0.1	0.0	0.1	0.2	2.2	3.8	4.7	3.8	2.2
No audio, informed										
AF-Next	0.0	0.0	0.0	0.0	0.0	22.4	33.4	43.1	29.4	21.0
Qwen2.5 7B	0.2	0.0	0.0	0.0	0.0	30.3	40.4	43.3	34.6	35.2
Qwen3 30B-A3B	0.0	0.0	0.0	0.0	0.0	31.9	46.7	45.1	38.7	20.4
Llama 3.1 8B	0.0	0.0	0.0	0.0	0.0	31.4	34.7	41.1	31.7	30.9
No audio, silent										
AF-Next	0.0	0.0	0.0	0.0	0.0	24.2	32.6	39.7	26.5	22.0
Qwen2.5 7B	0.0	0.0	0.0†	0.0	0.0	0.5	0.0	0.0	0.0	0.1
Qwen3 30B-A3B	0.0	0.0	0.0	0.0	0.0	0.7	0.1	0.2	0.7	0.3
Llama 3.1 8B	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.2	0.0	0.0

Table 14: BEANS-Next in-context learning controls and parse rates; inputs and formatting as in Table 11. (a) Gibbons and DCASE use macro-F1; other tasks use accuracy, all on a 0–1 scale. Zebra is an additional matching task not shown in Table 6. NatureLM-audio v1.0 does not accept multiple audio inputs. (b) Parse rate (%) for categorical answers, pooled over attempted examples: T1 MCQ ($n = 3{,}468$, three tasks), T2 behavior (3,657), T3 MCQ (5,807, seven), T4 matching (958, four), call type (2,000), and T4 detection (4,026, two). A parseable choice or label set need not be correct; an explicit no-detection answer counts where allowed. Binary presence and captioning are excluded. –: not applicable.

(a) Tier 4: in-context learning

	Few-shot task					
Model	Gibbons	Otters	DCASE	Crows	Unseen	Zebra
Real audio						
AF-Next	0.028	0.174	0.141	0.225	0.046	0.300
Qwen3-Omni	0.196	0.492	0.323	0.580	0.486	0.200
Gaussian noise						
AF-Next	0.000	0.286	0.000	0.250	0.050	0.275
Qwen3-Omni	0.184	0.142	0.094	0.065	0.018	0.275
No audio, informed						
AF-Next	0.000	0.250	0.000	0.250	0.252	0.250
Qwen2.5 7B	0.199	0.250	0.061	0.250	0.252	0.250
Qwen3 30B-A3B	0.246	0.250	0.200	0.250	0.252	0.250
Llama 3.1 8B	0.000	0.000	0.000	0.000	0.000	0.000
No audio, silent						
AF-Next	0.186	0.250	0.244	0.250	0.248	0.250
Qwen2.5 7B	0.000	0.000	0.000	0.000	0.000	0.000
Qwen3 30B-A3B	0.000	0.000	0.000	0.000	0.000	0.000
Llama 3.1 8B	0.000	0.000	0.000	0.000	0.000	0.000

(b) Parse rate for categorical answers

Model	T1 MCQ	T2 behavior	T3 MCQ	T4 matching	Call type	T4 detection
Real audio						
AF-Next	98	>99	91	69	68	96
Qwen3-Omni	54	98	56	88	100	82
NatureLM-audio v1.0	97	93	<1	–	>99	–
Gaussian noise						
AF-Next	99	>99	84	83	100	49
Qwen3-Omni	81	>99	54	49	100	24
NatureLM-audio v1.0	97	>99	4	–	100	–
No audio, informed						
AF-Next	99	>99	85	100	100	100
Qwen2.5 7B	>99	>99	92	100	100	100
Qwen3 30B-A3B	>99	>99	98	100	100	100
Llama 3.1 8B	>99	99	93	0	100	0
No audio, silent						
AF-Next	99	>99	85	100	100	100
Qwen2.5 7B	2	<1	1	0	0	0
Qwen3 30B-A3B	6	8	2	0	0	0
Llama 3.1 8B	1	<1	<1	0	0	0

F Examples

Table 15: BEANS-Next benchmark examples. Audio placeholders are omitted or compacted; each row shows the split/task family, prompt, and reference output used for evaluation.

Split / task family	Format	Prompt	Reference output
T1 vocalization description	MCQ	Which description best matches the vocalization? Options include a slow trill, low drone, sharp whistle, and staccato notes with nasal rise.	a series of rapid, repetitive, staccato notes followed by a nasal, rising inflection
T1 vocalization description	MCQ	Which description best matches the sound? Options include impulsive bursts, a melodic whistle, sharp staccato notes, and a low nasal moan.	A series of dense, impulsive bursts followed by a quiet, rapid trill and punctuated by loud clicks
T1 acoustic caption	Caption	Describe the acoustic character of this vocalization, including its frequency range.	The vocalizations consist of high-pitched, whistling one-note sounds spanning roughly 2,500 to 7,000 Hz, as well as buzzing, nasal one-note sounds ranging from 2,000 to 18,000 Hz.
T1 acoustic caption	Caption	Describe the acoustic character of this vocalization, including its recording quality.	This vocalization consists of a series of rapid rattles that increase in frequency, concluding with a single prolonged rattle. The sound is difficult to hear over background noise but becomes more audible towards the end of the recording.
T2 behavior	MCQ	Which acoustic behavior is the Flame-crested Manakin performing? Options: wing snap, no vocalization/silent, song.	Wing snap
T2 behavior	MCQ	Which vocalizations can be heard from the blue duck? Options: rapid drumming, continuous song, male whistle followed by female growl.	A male whistle followed by a female growl
T2 semantic caption	Caption	What can you hear in this recording?	A pair of Bronze-winged Ducks produces distinct vocalizations. The female emits quacking calls, while the male provides accompanying wheezy whistles.
T2 semantic caption	Caption	Describe the vocalization in this recording.	A Scarlet-and-white Tanager pair performs a duet, consisting of a single note from the male and a fine rattle from the female. Faint background vocalizations from other species are also audible.
T2 semantic caption	Caption	What can be heard in this recording?	Multiple Tawny Owls engage in a complex vocal interaction consisting of compound hooting and "ku-wick" calls. The calls include high-pitched, harsh hooting from a female responding to at least three males.

Continued on next page

Table 15 (continued)

Split / task family	Format	Prompt	Reference output
T3 structural caption	Caption	Describe the pattern of vocalizations in this recording.	The recording features a sequence of low-frequency calls from *Streptopelia decaocto*, which appear at the beginning, middle, and end of the clip. These calls are interspersed with higher-pitched vocalizations from *Sturnus unicolor* and *Luscinia megarhynchos*.
T3 highest pitch	MCQ	Which species produces the highest-pitched vocalization? Options: *Charadrius hiaticula*, *Numenius arquata*, *Motacilla alba*, *Erithacus rubecula*.	*Motacilla alba*
T3 longest vocalization	MCQ	Which species produces the longest vocalization? Options: *Poecile atricapillus*, *Pipilo erythrophthalmus*, *Cardinalis cardinalis*, *Geothlypis trichas*.	*Cardinalis cardinalis*
T3 co-occurrence	Binary	Are there any moments where both *Vireo olivaceus* and *Setophaga ruticilla* are calling simultaneously?	Yes
T3 species listing	Open-ended	How many different species are vocalizing, and what are they? Give scientific names.	4: *Corvus brachyrhynchos*, *Pipilo erythrophthalmus*, *Seiurus aurocapilla*, *Setophaga virens*
T3 per-species counting	Open-ended	How many calls from each species can you hear? Give scientific names.	*Pipilo erythrophthalmus*: 6; *Cardinalis cardinalis*: 5; *Geothlypis trichas*: 4; *Polioptila caerulea*: 2
T3 ordered summary	Open-ended	Describe each species in order of appearance, with call counts and frequency ranges.	*Corvus brachyrhynchos*: 3 calls, 920–2130 Hz; *Seiurus aurocapilla*: 2 calls, 2730–9550 Hz; *Regulus calendula*: 2 calls, 1430–8030 Hz; *Poecile atricapillus*: 6 calls, 2850–8830 Hz.
T4 gibbon few-shot detection	Multi-audio	Given three support sounds A–C, which are present in the query recording, if any?	Multiple pulse gibbon call
T4 giant otter call matching	Multi-audio	Given four call-type audio exemplars A–D, which call type best matches the query recording?	Begging Scream
T4 held-out species matching	Multi-audio	Given four species audio exemplars A–D, which species best matches the query recording?	*Minervarya agricola*