NATURELM-AUDIO: AN AUDIO-LANGUAGE FOUNDATION MODEL FOR BIOACOUSTICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) prompted with text and audio represent the state of the art in various auditory tasks, including speech, music, and general audio, showing emergent abilities on unseen tasks. However, these capabilities have yet to be fully demonstrated in bioacoustics tasks, such as detecting animal vocalizations in large recordings, classifying rare and endangered species, and labeling context and behavior—tasks that are crucial for conservation, biodiversity monitoring, and the study of animal behavior. In this work, we present NatureLM-audio, the first audio-language foundation model specifically designed for bioacoustics. Our carefully curated training dataset comprises text-audio pairs spanning a diverse range of bioacoustics, speech, and music data, designed to address the challenges posed by limited annotated datasets in the field. We demonstrate successful transfer of learned representations from music and speech to bioacoustics, and our model shows promising generalization to unseen taxa and tasks. Importantly, we test NatureLM-audio on a novel benchmark (BEANS-Zero) and it sets the new state of the art (SotA) on several bioacoustics tasks, including zero-shot classification of unseen species. To advance bioacoustics research, we also open-source the code for generating training and benchmark data, as well as for training the model

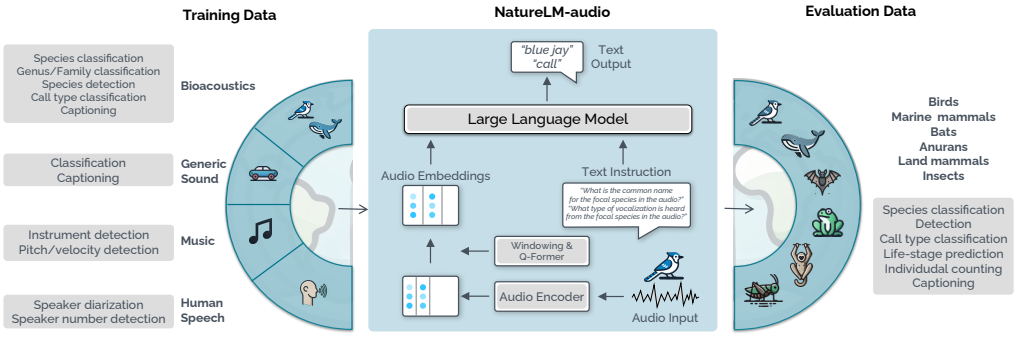

Figure 1: Overview of NatureLM-audio

1 INTRODUCTION

Bioacoustics, the study of sound production and reception in animals, aims to understand animal behavior (Fischer et al., 2013), monitor biodiversity (Stowell, 2022), and model the mechanisms of sound production and reception used in animal communication (Bradbury & Vehrencamp, 1998). It plays a vital role in conservation and ecological research, as animal vocalizations provide critical insights into ecosystem health, species interactions, and population dynamics. By enabling the detection of endangered species and tracking migration patterns, bioacoustic research directly contributes to biodiversity monitoring and conservation efforts (Rutz et al., 2023; Stevens et al., 2024).

In recent years, machine learning has taken on an increasingly pivotal role in bioacoustic research. Beyond its applications in large-scale ecological monitoring, machine learning has also opened up new frontiers in the study of animal communication, enabling discoveries like the ability of marmosets (Oren et al., 2024), dolphins (King & Janik, 2013), and elephants (Pardo et al., 2024) to use specialized vocalizations to label their conspecifics. Yet, because of obvious data collection and annotation difficulties, these studies often rely on strongly labeled small datasets (Stowell, 2022) and thus require careful statistical analysis to measure the significance of results and avoid over-fitting. At the same time, large volumes of unannotated bioacoustics data are recorded daily, particularly through passive acoustic monitoring (PAM, Dufourq et al. (2021)) or citizen science platforms e.g. Xeno-canto (Vellinga & Planqué, 2015)). There is thus a growing need for machine learning tools capable of performing tasks such as detection, classification, and annotation on these data at scale. The recent successes of large scale artificial intelligence models in various domains (e.g. natural language processing, vision, games) also point to the possibility of leveraging these huge volumes of raw data to learn accurate and generalizable representations of bioacoustics signals (Ghani et al., 2023; Boudiaf et al., 2023).

Existing bioacoustics machine learning models are typically designed for specific species or tasks (Dufourq et al., 2021; Kahl et al., 2021; Cauzinille et al., 2024), showing limited generalizability beyond their predefined scope. Many traditional studies rely on small datasets focused on a few species and individuals, validating results through statistical measures despite over-fitting risks. Newer models such as BirdNET (Kahl et al., 2021) and Perch (Ghani et al., 2023) perform well in specific tasks such as bird classification but require training of a classifier specific to each target taxa. Instead, we propose a single foundation model that works across taxa. Recently, self-supervised and audio-language contrastive models, AVES (Hagiwara, 2023) and BioLingual (Robinson et al., 2024), have exhibited notable results on bioacoustics benchmarks, though they remain constrained by their training paradigms (discriminative and contrastive, respectively), which restrict the range of tasks they can address.

In recent years, foundation models, which learn patterns in large amounts of broad data (generally via self-supervision), have shown promising performance across a wide range of tasks (Bommasani et al., 2021). While transformer-based large language models (LLMs) are currently the most prominent examples, other architectures, such as diffusion models (Kingma et al., 2021), are also emerging as foundation models in some domains. These models' ability to handle unseen tasks, perform in-context learning, and respond to prompts positions them as a compelling alternative to traditional machine learning methods, which often rely on laboriously annotated data, expensive computational resources, and often-lacking machine learning expertise.

While multimodal large language models (LLMs), particularly vision-language models (VLMs), have been explored for biodiversity and conservation research (Miao et al., 2024), there is relatively little effort dedicated to building and investigating large audio-language models (LALMs) for bioacoustics. LALMs have shown significant promise in processing human speech (Rubenstein et al., 2023; Wang et al., 2024; Wu et al., 2023a; Zhang et al., 2024), music (Gardner et al., 2023; Agostinelli et al., 2023), and general audio tasks (Tang et al., 2024; Chu et al., 2024; Gong et al., 2023), and they hold the potential to bring transformative advancements to bioacoustics as well.

In this paper, we present NatureLM-audio, an audio-language foundation model specifically designed for bioacoustics tasks, including classification, detection, and captioning. To the best of our knowledge, NatureLM-audio is the first model of its kind. Inspired by the cross-taxa transfer observed in previous research, such as between human and gibbons (Cauzinille et al., 2024) and birds and whales (Ghani et al., 2023), we incorporate speech and music tasks into the training process. We show that representations learned from these domains successfully transfer to animal vocalizations, demonstrating generalization across species. Importantly, we augment an already existing animal sounds classification and detection benchmark, BEANS (Hagiwara et al., 2023), with additional tasks such as call-type prediction, lifestage classification, captioning, and individual counting. With these, we test cross-domain learning capabilities of the model and zero-shot transfer to unseen taxa and tasks. We name this new benchmark BEANS-Zero. Unlike existing bioacoustics benchmarks such as Perch (Ghani et al. (2023) for bird detection) and BirdSet (Rauch et al. (2024) for bird classification), we do not focus solely on birds and we go beyond species classification. Additionally, the prompts and the audio are described in natural language in our dataset. This has the potential to accelerate the research in LALMs.

Our contributions are thus as follows:

- **Model**: We introduce NatureLM-audio, to the best of our knowledge, the first audio-language foundation model for bioacoustics with carefully curated training datasets comprising of animal vocalization, human speech, and music.
- **Domain transfer** We show that the model transfers beyond the species originally trained on and demonstrate its zero-shot capability on unseen taxa and species.
- **Task transfer** We test our model on a novel benchmark (BEANS-Zero) that goes beyond species classification and even includes a completely unseen task (individual counting). For the first time, we show positive transfer from speech and music data to bioacoustics tasks.

2 Related Work

Most prior work on audio-language models has focused on human speech processing. For example, models like SpeechGPT (Zhang et al., 2023), Speech-LLaMA (Wu et al., 2023a), AudioLM (Borsos et al., 2023), AudioPaLM (Rubenstein et al., 2023), AudioGPT (Huang et al., 2023), SpiRit-LM (Nguyen et al., 2024), and SpeechLM (Zhang et al., 2024) mostly focus on building language models that can perceive and produce human speech. Such models may be fine-tuned for downstream bioacoustics tasks requiring expensive computational resources and expertise. Instead, our model shows promising generalization to unseen species and tasks.

Recently, more generic language models with audio perception capabilities have been released. Pengi (Deshmukh et al., 2023) uses an audio encoder and a text encoder mapped onto an LLM to solve audio-to-text tasks. SALMONN (Tang et al., 2024) uses dual audio encoders and integrates Q-Former (Li et al., 2023) to improve the handling of speech and general audio inputs. Qwen-audio (Chu et al., 2023) adopts a multi-task learning approach with the introduction of the Speech Recognition with Timestamp (SRWT) task. LTU (Gong et al., 2023) builds an open-ended question-answer dataset and uses curriculum learning strategies to enhance its generalization capabilities. Similar multimodal language models have been proposed for music, such as MU-LLaMA (Liu et al., 2023) and LLark (Gardner et al., 2023). Recent foundation models such as AVES (Hagiwara, 2023) and BioLingual (Robinson et al., 2024) have exhibited notable results on bioacoustic tasks, although their training paradigms and architectures restrict the range of tasks they can address.

Although animal sounds and vocalizations are often part of generic audio datasets, such as AudioSet (Gemmeke et al., 2017) and audio caption datasets (Kim et al., 2019; Mei et al., 2023), these datasets are often too general and lack the fine-grained details necessary for tasks like species classification, behavior analysis, or monitoring in ecology and bioacoustics. As a consequence, LALMs trained on these datasets produce at best generic labels e.g., 'bird' and not the name of the species. We address this limitation by proposing an open multi-task diverse training set and a LALM, NatureLM-audio, that offers robust representations for bioacoustics.

While there are specific bioacoustics benchmarks like BIRB (Hamer et al., 2023) for bird vocalization retrieval and BEANS (Hagiwara et al., 2023) for classification/detection, the field of bioacoustics has yet to see the development of dedicated benchmarks similar to those in human speech and music, such as Dynamic-SUPERB (Huang et al., 2024) or AIR-Bench (Yang et al., 2024). This leaves a gap for advancing the evaluation of bioacoustics models, particularly in zero-shot learning and task generalization.

With this work, we aim to bridge these gaps by introducing NatureLM-audio, a model specifically designed for bioacoustics tasks, and enhancing bioacoustic benchmarks to assess cross-species and cross-task generalization, introducing BEANS-Zero.

3 Methods

3.1 Training Dataset Creation

To train an audio-text model for bioacoustics, we compile a diverse dataset of text-audio pairs (Table 1). The data is collected through a combination of prompting on existing audio datasets, creating

Task[a]	Dataset	# Hours	# Samples
CAP	WavCaps (Mei et al., 2023)	7568	402k
CAP	AudioCaps (Kim et al., 2019)	145	52k
CLS	NSynth (Engel et al., 2017)	442	300k
CLS	LibriSpeechD (Edwards et al., 2018)	156	16k
CLS, DET, CAP	Xeno-canto (Vellinga & Planqué, 2015)	10416	607k
CLS, DET, CAP	iNaturalist (iNaturalist)	1539	320k
CLS, DET, CAP	Watkins (Sayigh et al., 2016)	27	15k
CLS, DET	Animal Sound Archive (Museum für Naturkunde Berlin)	78	16k
DET	Xeno-canto-detection (Vellinga & Planqué, 2015)	2749	670k
DET	Sapsucker Woods (Kahl et al., 2022a)	285	342k
DET	Sierra Nevada (Kahl et al., 2022b)	61	22k
DET	University of Hawai'i at Hilo (Navine et al., 2022)	94	34k

Table 1: Training tasks and datasets [a] CLS: classification, DET: detection, CAP: captioning

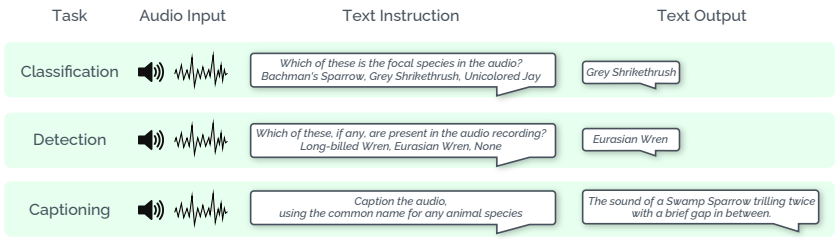

Figure 2: Examples of training instances

new LLM-generated text labels, and mixing new, procedurally-augmented audio data. The data is comprised of bioacoustic audio, general audio, speech, and music datasets. Figure 2 shows some examples of instances used for training NatureLM-audio.

3.1.1 BIOACOUSTIC DATA

Species Classification: We curate existing large-scale bioacoustic archives into a common format. We process Xeno-Canto (Xeno-canto), iNaturalist (iNaturalist), Animal Sound Archive (Museum für Naturkunde Berlin), and Watkins (all-cuts, Sayigh et al. (2016)) into a common format. Specifically, we handle differences in common name and scientific name across datasets by joining all datasets to the GBIF taxonomy backbone (GBIF Secretariat, 2023). We then prompt the model to predict either the scientific or common name of the focal species, or the scientific or common names of all species in the recording. This requires the model to generate the common name or scientific name of the species directly. In many cases, we may know an animal vocalization is one of a subset of species—for example, based on location. To allow for this, we also generate prompts with a set of options injected into the question. For thirty percent of prompts, we sample "random" negatives by selecting from all common names or scientific names in our dataset. For the remaining prompts, we randomly choose an ancestor level of either family, order, or phylum, and sample "hard" negatives with the same ancestor as the correct species. The number of negatives is chosen randomly from one up to a maximum of thirty-five. To avoid data leakage during evaluation, we exclude a set of held-out species as well as the cbi data used in BEANS-Zero.

Species Detection: We use the same datasets as for species classification, but prompt the model to ask whether the recording contains one of a set of options, or 'None'. Options are sampled in the same way as for classification, with a mix of random and hard negatives. In fifty percent of prompts, the correct species is not included in the set of options, with a correct answer of 'None'. We additionally prompt

To help bridge the gap between focal train recordings and noisy soundscape recordings common at inference, we also generate a noise-augmented detection training set from Xeno-canto. We use per-channel energy normalization (PCEN Lostanlen et al. (2018)) as a form of noise-gate for bird vocalization activity detection. Then, we separate each detected segment into four stems using the

4-stem Bird-MixIT source separation model (Denton et al., 2022). Because the separation model may over-separate sources and does not label stems with source names, we use the YAMNet model (Howard et al., 2017) trained on the AudioSet dataset (Gemmeke et al., 2017) to select solely the stems with high probability on the AudioSet animal classes (with ids between 67 to 131). Correspondingly, for each stem we take the maximum probability across the classes, we average the values across the time frames, and we sum the stems with values higher than 0.5.

Because Xeno-Canto comprises mostly focal recordings, we account for the covariate shift in soundscapes by adding noise—audio that does not contain animal vocalizations, speech, or music. The noise samples are extracted from the following datasets: boat engine sounds from ShipsEar (Santos-Domínguez et al., 2016), Deepship (Irfan et al., 2021) and Orcalab (Poupard et al., 2020), non-animal, non-music sound classes from FSD50K (Fonseca et al., 2021) and Urbansound (Salamon & Jacoby, 2014), and all the classes from TUT2016 (Mesaros et al., 2016), IDMT (Abeßer et al., 2021), Demand (Thiemann et al., 2013), and Wham noise (Wichern et al., 2019). The noise is added programmatically, using random files at a random signal-to-noise ratio (SNR) sampled from a uniform distribution ranging from -5dB to 10dB.

In addition, we used soundscape recording datasets from Sapsucker Woods (SSW, Kahl et al. (2022a)), Sierra Nevada (SNE, Kahl et al. (2022b)), and the University of Hawai'i at Hilo (UHH, Navine et al. (2022)) for detection tasks. Following the approach used in the detection datasets from BEANS, we split the audio into 10-second windows with a 5-second overlap, and treated it as a multi-label classification problem. Species with more than 100 occurrences were selected as target labels, while species with fewer occurrences were grouped into an "other" class.

Captioning: We use the AnimalSpeak (Robinson et al., 2024) dataset for bioacoustic captioning. AnimalSpeak combines bioacoustic datasets into a language-model-captioned audio-text dataset. However, due to scale, the large segment of AnimalSpeak from Xeno-Canto was not captioned with a language-model, and used only templated captions. We further process Xeno-Canto with Gemini-1.0-pro (Gemini Team, 2024) following the same method used to create AnimalSpeak, and use these LLM-generated captions in addition to the original captions.

Call-type and Lifestage: We include multiple new bioacoustic tasks which can be expressed based on the Xeno-Canto metadata. Specifically, predicting the life stage of birds, predicting call-types, and differentiating between calls and songs. Compared to species classification alone, included in existing datasets, the ability to perform these tasks at scale could significantly enhance the precision of ecological monitoring and behavior studies.

3.1.2 General Audio

We include WavCaps (Mei et al., 2023) and AudioCaps (Kim et al., 2019) for general audio captioning. We observe that, in the creation of WavCaps, some recordings originally had metadata relevant to bioacoustics and specific species. However, these were lost in the general-domain captioning, producing captions which are too generic for our purpose. We detect these cases by processing the original metadata, and re-process the metadata prompting Gemini-1.0-pro to produce bioacoustic captions. We include these new bioacoustic captions in addition to the original captions.

3.1.3 Music

Pitch, timbre qualities of animal vocalizations, the number of animals in a recording are often key acoustic features used by biologists to classify context and behavior. We use NSynth 2.3.3 (Engel et al., 2017) to create a set of tasks that may help bioacoustics downstream tasks. We generate text prompts for *pitch detection* in Hz, *instrument name*, and *velocity*, ranging 0 to 1. Additionally, we use the timbre 'qualities' labels to create *text descriptions* for each audio. For instance, if the sound is 'distorted,' we generate descriptions such as 'This sound has a distinctive crunchy sound and presence of many harmonics.' or 'This sound is distorted'. Moreover, we create synthetic mixtures by layering one to three different instruments. In this case we generate, two task: predicting the *number of instruments* and identifying the *instrument names*.

Task[a]	Dataset	Description	# Size[b]	# Labels (type)
CLS	esc50	generic sound	400	50 (sound type)
CLS	watkins	marine mammals	339	31 (species)
CLS	cbi	birds	3620	264 (species)
CLS	humbugdb	mosquito	1859	14 (species)
DET	dcase	birds & mammals	13688	20 (species)
DET	enabirds	birds	4543	34 (species)
DET	hiceas	cetaceans	1485	1 (species)
DET	rfcx	birds & frogs	10406	24 (species)
DET	gibbons	gibbons	18560	3 (call type)
CLS	unseen-cmn	birds etc.	931	300 (species)
CLS	unseen-sci	birds etc.	931	300 (species)
CLS	lifestage	birds	493	3 (stage)
CLS	call-type	birds	15439	2 (call/song)
CAP	captioning	birds etc.	29002	(open-ended)
CLS	zf-indv	zebra finches	2346	4 (# of indv.)

Table 2: Evaluation tasks and datasets of BEANS-Zero. [a] CLS: classification, DET: detection, CAP: captioning. [b] The numbers of samples for classification and captioning, and the number of 5-second "chunks" for detection (see Section 3 for more details)

3.1.4 SPEECH

We use the speech diarization dataset based on LibriSpeech (Edwards et al., 2018), which contains synthetic mixtures of two or three speakers. We use this to derive the *number of speakers* task, which we believe has interesting applications for monitoring individuals if transferred to bioacoustics.

3.2 EVALUATION DATA: THE BEANS-ZERO BENCHMARK

One contribution of this work is a new benchmark for bioacoustics: BEANS-Zero (Table 2). With BEANS-Zero, we go beyond traditional species classification, introducing tasks such as call-type prediction, lifestage classification, captioning, and individual counting (which is not seen during training). To build this set of tasks, we first used the test portion of the benchmark BEANS (Hagiwara et al., 2023) for evaluating our models on common bioacoustics datasets and tasks, which include:

- esc50 (Piczak, 2015): Generic environmental audio classification with 50 labels.
- watkins (Sayigh et al., 2016): Marine mammal species classification with 31 species.
- cbi (Howard et al., 2020) Bird species classification with 264 labels from the Cornell Bird Identification competition hosted on Kaggle.
- humdubdb (Kiskin et al., 2021) Mosquito wingbeat sound classification into 14 species.
- dcase (Morfi et al., 2021) Mammal and bird detection from DCASE 2021 Task 5: Few-shot Bioacoustic Event Detection (20 species)
- enabirds (Chronister et al., 2021) Bird dawn chorus detection with 34 labels.
- hiceas (Center, 2022) Minke whale detection from the Hawaiian Islands Cetacean and Ecosystem Assessment Survey (HICEAS) (1 label).
- rfcx (LeBien et al., 2020): Bird and frog detection from the Rainforest Connection (RFCx) data with 24 species.
- gibbons (Dufourq et al., 2021): Hainan gibbon detection with 3 call type labels.

We also include novel bioacoustics datasets including:

- unseen-cmn: 300 species held out from AnimalSpeak (Robinson et al., 2024) with common (English) names. For a dataset of medium difficulty, we hold out species at random whose genus is reasonably well-represented in the training set (at least 100 recordings.)
- unseen-sci: same recordings as above, but predicted with scientific (Latin) names

- `lifestage`: Predicting the lifestage of birds across many species. Newly curated from Xeno-canto (Xeno-canto).
- `call-type`: Classifying song-vs. call across multiple species of birds. Newly curated from Xeno-canto (Xeno-canto).
- `captioning`: Captioning bioacoustic audio on AnimalSpeak (Robinson et al., 2024)
- `zf-indv` (Elie & Theunissen, 2016): Counting the number of zebra finch individuals

Some of these tasks, in particular captioning of bioacoustic audio, are previously unstudied. Captioning allows for automatic generation of descriptive annotations of animal sounds, enhancing our understanding of species behaviors and communication patterns. Improvements in other new tasks, such as cross-species lifestage and call-type prediction, would allow finer-grained ecological monitoring at scale.

For evaluation, we use accuracy for classification, macro-averaged F1 for detection, and SPIDEr (Liu et al., 2017) for captioning. We opt for F1 instead of mean average precision (mAP), which is originally used in BEANS for detection, as F1 is better suited for generative models, whereas mAP assumes a smooth ranking of candidates, which is less appropriate for evaluating generative tasks.

3.3 NatureLM-audio Architecture

Our model follows a generic audio-to-text architecture used in prior works, such as SALMONN (Tang et al., 2024), Qwen2-audio (Chu et al., 2024), and LTU (Gong et al., 2023), which are large audio-language models trained on paired audio-text data for tasks including speech, music, and general audio events. Figure 1 provides an overview of the NatureLM-audio architecture.

Specifically, NatureLM-audio first encodes the audio input via an audio encoder, in this case BEATs (Chen et al., 2023), which has achieved SotA on multiple audio tasks. To connect the BEATs embeddings with the LLM we use a Q-Former (Li et al., 2023) applied at the window level as proposed in SALMONN (Tang et al., 2024). Similarly to the existing LALMS we use an LLM to produce text, in this case Llamma 3.1-8b (Dubey et al., 2024), which is fine-tuned with LoRA (Hu et al., 2022). The parameters of the LLM (except for the adapter layers) remain frozen during training, while the audio encoder and Q-Former are unfrozen. The model takes an audio $\boldsymbol{a}$ and an instruction $\boldsymbol{x}$ as its input, and produces a text sequence $\boldsymbol{x}_{<t}$ as the output. The model is trained under the loss function:

$$\boldsymbol{h} = f_W(\text{Encoder}(\boldsymbol{a})) \tag{1}$$

$$\boldsymbol{z} = p_\varphi^Q(\boldsymbol{q}, \boldsymbol{h}) \tag{2}$$

$$L = -\sum \log p_\theta^{LM}(\boldsymbol{x}_{<t} | \boldsymbol{z}, \boldsymbol{x}) \tag{3}$$

where Encoder is the pretrained BEATs (Chen et al., 2023) audio encoder, f_W is a function that converts consecutive W audio frames into a window, p_φ^Q is the Q-Former model with trainable parameters φ that converts a window into a sequence of text representations $\boldsymbol{z}$ using query $\boldsymbol{q}$, and p_θ^{LM} is the pretrained LLM with trainable parameters θ.

3.4 Training Method

Our training method is heavily motivated by curriculum learning (Soviany et al., 2021) where machine learning algorithms start with simpler, easy to learn instances and gradually shift to more difficult ones, as done in other audio foundation models (Tang et al., 2024; Gong et al., 2023). We train in the following two stages:

- Stage 1 (Perception Pretraining): We pretrain the model exclusively on the task of focal species classification, classifying vocalizations of thousands of animal species. Species classification is highly deterministic, allowing opportunity to learn a robust connection between language and audio. We also choose to train on this task individually as it is foundational to other tasks in bioacoustics.
- Stage 2 (Generalization Fine-tuning): In the second stage, we introduce a variety of bioacoustic and other tasks that build on the robust classification performance of the first stage.

Model	esc50	watkins	cbi	humbugdb	dcase	enabirds	hiceas	rfcx	gibbons
LLM w/o audio	0.020	0.041	0.005	0.073	0.000	0.001	0.210	0.000	<u>0.013</u>
SALMONN	0.320	0.041	0.004	**0.090**	0.005	0.004	0.097	0.002	0.005
Qwen2-audio	0.307	0.041	0.004	0.070	0.005	0.004	0.097	0.002	0.005
BioLingual	<u>0.600</u>	<u>0.257</u>	<u>0.705</u>	<u>0.085</u>	<u>0.036</u>	<u>0.109</u>	**0.429**	<u>0.004</u>	**0.018**
NatureLM-audio	**0.635**	**0.646**	**0.755**	0.073	**0.052**	**0.279**	<u>0.390</u>	**0.039**	0.003

Table 3: Main zero-shot results on BEANS-Zero. We used accuracy for classification, and F1 for detection tasks. The best and the second best metrics are highlighted and underlined per each dataset

This includes detection, captioning, lifestage prediction, and call-type prediction. We also include speech and music data in this second stage, hoping to transfer to bioacoustic tasks.

We trained from scratch (i.e., random initialization of the Q-Former and LoRA) rather than fine-tuning existing models or checkpoints, such as SALMONN's. This allows for more flexibility in terms of choosing the latest LLM, with the most knowledge of animal species, and the most relevant architectural components (e.g. excluding memory-heavy parts of current LALMs such as the speech encoder Whisper (Radford et al., 2022)).

4 Experiments

4.1 Training and Evaluation Details

We train our model on the full curated training set (Section 3.1). To assess the model's generalization we created hold-out splits for Xeno-canto, iNaturalist, Animal Sound Archive, and Watkins datasets, used solely for benchmarking.

We initialize the audio encoder weights using an existing BEATs checkpoint [1] and fully fine-tune it. For Llama, we start from Llama-3.1-8B-Instruct and fine-tune all attention layers with LoRA (rank: 64, alpha: 64, dropout: 0.1).

We train with our two proposed stages. In both stages we use a linear warmup, cosine scheduler, peak learning rate of 9.0×10^{-5}, and a batch size of 64. We decode using beam search with two beams, a repetition penalty of 1.0, and a length penalty of 1.0.

We consider several inference methods depending on the task type. Species-classification tasks involve single-label prediction: we prompt the model to output the species name from the recording. To handle the case where the LLM outputs text which is not an allowed label, we match to the closest label according to the Levenshtein distance. We choose the Levenshtein distance for its simplicity and because species names, in particular Latin names, have high character-overlap with related names. We note this may not be optimal for general audio classification.

For multilabel detection tasks, we range from detecting a large number of species to only a single species, depending on the dataset. When detecting only a few species (ten or less), we include the options in the prompt. Otherwise we prompt the model to predict all species in the audio window, if any. In both cases, the model outputs all detected species, or 'None'. We discard detections with low character-overlap with the allowed labels.

Our baselines include CLAP-like models (Wu et al., 2023b), which cannot naively perform multilabel detection. To handle this, we create a negative "template" for each detection task, as proposed in (Miao et al., 2023). We consider each label a detection positive for CLAP if the audio is more similar to the label than to the negative template in the CLAP model's embedding space.

4.2 Species Classification and Detection

Table 3 shows the main results measured on the BEANS-Zero species classification and detection datasets. Our baselines include an LLM (the original Llama-3.1-8B-Instruct model without fine-tuning, Dubey et al. (2024)) without audio input, SALMONN (Tang et al., 2024), BioLingual (Robinson et al., 2024), and Qwen2-audio (Chu et al., 2024). All baselines are evaluated in

[1] `BEATs_iter3_plus_AS2M_finetuned_on_AS2M_cpt2.pt`

	cbi	dcase-bird	enabirds
BirdNET	0.609	0.035	0.490
Perch	0.744	0.035	0.164
NatureLM-audio	0.755	0.088	0.279

Table 4: Comparison with bird vocalization models

	unseen-cmn[a]	unseen-sci[b]
Supervised SotA	0.547	0.614
NatureLM-audio	0.116	0.196
baseline (CLAP)	0.034	0.004

Table 5: Generalization to unseen species in terms of classification accuracy for: [a] common (English) names and [b] latin/scientific names

the same way as NatureLM-audio. As shown in the table, the outputs from the LLM without audio input, SALMONN, and Qwen2-audio are largely random on the bioacoustic datasets, failing to properly interpret the input audio or follow the instructions. In contrast, NatureLM-audio achieved state-of-the-art zero-shot performance on 6 out of 9 datasets, and delivered competitive results on the remaining tasks from the BEANS-Zero benchmark. We observe that for some of those three remaining tasks, our current training data contains little signal, for example on `humbugdb` (Kiskin et al., 2021) which classifies species by mosquito wingbeat sounds not generated by a vocal tract. We also note that performance of baselines on the general audio auxiliary dataset ESC50 (Piczak, 2015) may be reduced by the use of the Levenshtein distance, as our pipeline is optimized for bioacoustic tasks.

We also compared NatureLM-audio with bird vocalization classification models, namely BirdNET (Kahl et al., 2021) and Perch (Ghani et al., 2023), to evaluate the zero-shot capabilities of our model. We compare on the subset of BEANS-Zero classifying or detecting exclusively bird species, plus the portion of DCASE with bird species. The results are presented in Table 4. Since both BirdNET and Perch were trained in a supervised manner on datasets that significantly overlap with our bird evaluation datasets, this is not a fully fair comparison, and their performance should be considered as topline results. Nevertheless, our model demonstrated strong zero-shot bird vocalization classification capabilities. In particular, we achieve a new SotA for the cbi dataset, classifying vocalizations of hundreds of birds, and achieve competitive results with the bird-specific models on both detection tasks.

4.3 Generalizing to Unseen Species

We further evaluate the model's ability to generalize to completely unseen taxa using the newly added datasets in BEANS-Zero. They consist of recordings of held-out species from Xeno-canto, iNaturalist, Animal Sound Archive, and Watkins. As a topline, we compare against BioLingual, which has seen these species in training and serves only as an indicator of fully supervised classification performance. As baselines, we consider a theoretical random baseline of 0.3% accuracy (with 300 classes, random chance yields an accuracy of $\frac{1}{300} \approx 0.3\%$) and CLAP-LAION (Elizalde et al., 2023), a general-domain audio model which, similar to our model, is unlikely to have seen these species during training. We compare the performance when predicting common as well as scientific names.

Our model significantly outperforms the random baseline, demonstrating generalization to completely unseen species. Specifically, on the unseen species test set, our model achieves an accuracy of 19.6%, which is substantially higher than the random baseline of 0.3%. This indicates that the model has learned generalizable features that extend beyond the species it was trained on. Additionally, our model outperforms the CLAP-LAION baseline, further emphasizing its ability to generalize. Our model in particular excels when predicting with scientific (Latin) names (*unseen-sci*), which have consistent hierarchical structure it may learn to exploit.

	lifestage	call-type	captioning	zf-indv
SotA	0.676	0.499	0.009	0.225
NatureLM-audio	0.763	0.810	0.494	0.383

Table 6: Results on BEANS-Zero novel bioacoustics tasks. We report accuracy for classification, and SPIDEr (Sharif et al., 2018) for captioning. SotA is SALMONN for captioning and Biolingual for the remaining tasks.

4.4 Novel Bioacoustic Tasks

We evaluate the model's abilities beyond species prediction with several bioacoustic tasks newly added to BEANS-Zero, which have, to the best of our knowledge, not been studied at a cross-species level. We additionally include `zf-indv`, a completely unseen task counting the number zebra finches in a recording (Elie & Theunissen, 2016). We compare against BioLingual (Robinson et al., 2024) for discriminative tasks and SALMONN (Tang et al., 2024) for captioning. On each of these tasks, our model sets the state-of-the-art.

4.5 Ablation on Speech and Music

To investigate the impact of speech and music on downstream task performance, we run an ablation on stage-2 training with and without speech and music data. We train both stage-2 models for 200k steps, and evaluate their ability to perform the unseen task of counting zebra-finch individuals in a recording. The model trained with speech scores .379 on this task, similar to our full model. The model trained without speech scores an accuracy of .243, approximately random, and qualitatively predicts a single speaker for every recording. This result suggests the ability to count vocalizing birds transfers from human speech and music, for which our training data includes counting human speakers in a recording. We include the ablation performance on all tasks in the Appendix, as shown in Tables 8 and 9.

5 Conclusion

We presented NatureLM-audio, the first audio-language foundation model specifically designed for bioacoustics, demonstrating its potential to address critical tasks such as classifying and detecting animal vocalizations, and decoding context, call types, and individuals across species. By leveraging a carefully curated dataset spanning bioacoustics, speech, and music data, NatureLM-audio sets the new state-of-the-art on multiple tasks, including zero-shot classification of unseen species. Moreover, our model demonstrates positive transfer across both domain and tasks, performing well on a novel benchmark (BEANS-Zero), which includes new bioacoustic tasks such as captioning and individual counting. To further accelerate research and the development of more robust models in the field, we have open-sourced the code for generating both training and benchmarking data.

We plan to extend this work by incorporating more diverse tasks and datasets, improving the text-based LLM backbone with bioacoustic-specific texts, and enhancing the model's multilingual capabilities. Additionally, we aim to introduce new modalities, such as motion and image data, leading to models like NatureLM-motion and NatureLM-image. Lastly, we will explore the model's generative abilities, enabling it to produce audio tokens for tasks such as animal sound generation and audio denoising.

While NatureLM-audio offers significant potential for advancing biodiversity monitoring and conservation, several ethical concerns must be addressed. First, there is a potential bias towards bird vocalizations due to the overrepresentation of bird datasets, which could limit the model's effectiveness in other domains. Second, the model's ability to detect and classify endangered species could be misused for illegal activities such as poaching, posing a threat to wildlife. Finally, unintended consequences on animal behavior and ecology must be considered, particularly when deploying LLMs, known for their issues including hallucinations and biases (Kuan et al., 2024). These systems may interfere with the behavior of the species being studied, and the long-term ecological impact of widespread passive monitoring is still unknown. Careful deployment and responsible use are essential to mitigate these risks.

REFERENCES

Jakob Abeßer, Saichand Gourishetti, András Kátai, Tobias Clauß, Prachi Sharma, and Judith Liebetrau. Idmt-traffic: an open benchmark dataset for acoustic traffic monitoring research. In *2021 29th European Signal Processing Conference (EUSIPCO)*, pp. 551–555. IEEE, 2021.

Andrea Agostinelli, Timo I. Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, Matt Sharifi, Neil Zeghidour, and Christian Frank. MusicLM: Generating music from text, 2023. URL https://arxiv.org/abs/2301.11325.

Rishi Bommasani et al. On the opportunities and risks of foundation models. *ArXiv*, abs/2108.07258, 2021. URL https://arxiv.org/abs/2108.07258.

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, and Neil Zeghidour. AudioLM: a language modeling approach to audio generation, 2023. URL https://arxiv.org/abs/2209.03143.

Malik Boudiaf, Tom Denton, Bart Van Merriënboer, Vincent Dumoulin, and Eleni Triantafillou. In search for a generalizable method for source free domain adaptation. In *International Conference on Machine Learning*, pp. 2914–2931. PMLR, 2023.

Jack W. Bradbury and Sandra L. Vehrencamp. *Principles of animal communication*, volume 132. Sinauer Associates Sunderland, MA, 1998.

Jules Cauzinille, Benoît Favre, Ricard Marxer, Dena Clink, Abdul Hamid Ahmad, and Arnaud Rey. Investigating self-supervised speech models' ability to classify animal vocalizations: The case of gibbon's vocal identity. In *Interspeech*. ISCA, 2024.

NOAA Pacific Islands Fisheries Science Center. Hawaiian islands cetacean and ecosystem assessment survey (HICEAS) towed array data. *Edited and annotated for the 9th International Workshop on Detection, Classification, Localization, and Density Estimation of Marine Mammals Using Passive Acoustics (DCLDE 2022)*, 2022. doi: https://doi.org/10.25921/e12p-gj65.

Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, Wanxiang Che, Xiangzhan Yu, and Furu Wei. BEATs: Audio pre-training with acoustic tokenizers. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 5178–5193. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/chen23ag.html.

Lauren M. Chronister, Tessa A. Rhinehart, Aidan Place, and Justin Kitzes. An annotated set of audio recordings of Eastern North American birds containing frequency, time, and species information. *Ecology*, 102(6):e03329, 2021. doi: https://doi.org/10.1002/ecy.3329. URL https://esajournals.onlinelibrary.wiley.com/doi/abs/10.1002/ecy.3329.

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

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen2-audio technical report, 2024. URL https://arxiv.org/abs/2407.10759.

Tom Denton, Scott Wisdom, and John R Hershey. Improving bird classification with unsupervised sound separation. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 636–640. IEEE, 2022.

Soham Deshmukh, Benjamin Elizalde, Rita Singh, and Huaming Wang. Pengi: An audio language model for audio tasks. In *Advances in Neural Information Processing Systems*, volume 36, pp. 18090–18108, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/3a2e5889b4bbef997ddb13b55d5acf77-Paper-Conference.pdf.

Abhimanyu Dubey et al. The Llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Emmanuel Dufourq, Ian Durbach, James P. Hansford, Amanda Hoepfner, Heidi Ma, Jessica V. Bryant, Christina S. Stender, Wenyong Li, Zhiwei Liu, Qing Chen, Zhaoli Zhou, and Samuel T. Turvey. Automated detection of Hainan gibbon calls for passive acoustic monitoring. *Remote Sensing in Ecology and Conservation*, 7(3):475–487, 2021. doi: https://doi.org/10.1002/rse2.201. URL https://zslpublications.onlinelibrary.wiley.com/doi/abs/10.1002/rse2.201.

Erik Edwards, Michael Brenndoerfer, Amanda Robinson, Najmeh Sadoughi, Greg P Finley, Maxim Korenevsky, Nico Axtmann, Mark Miller, and David Suendermann-Oeft. A free synthetic corpus for speaker diarization research. In *International Conference on Speech and Computer*, pp. 113–122. Springer, 2018.

Julie E Elie and Frederic E Theunissen. The vocal repertoire of the domesticated zebra finch: a data-driven approach to decipher the information-bearing acoustic features of communication signals. *Animal cognition*, 19:285–315, 2016.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. CLAP: Learning audio concepts from natural language supervision. In *ICASSP*, pp. 1–5, 2023. ISBN 978-1-7281-6327-7. doi: 10.1109/ICASSP49357.2023.10095889.

Jesse Engel, Cinjon Resnick, Adam Roberts, Sander Dieleman, Mohammad Norouzi, Douglas Eck, and Karen Simonyan. Neural audio synthesis of musical notes with WaveNet autoencoders. In *International Conference on Machine Learning*, pp. 1068–1077. PMLR, 2017.

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.

E Fonseca, X Favory, J Pons, F Font, and X Serra. Fsd50k: an open dataset of human-labeled sound events. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:829–852, 2021.

Josh Gardner, Simon Durand, Daniel Stoller, and Rachel M Bittner. LLark: A multimodal instruction-following language model for music. *arXiv preprint arXiv:2310.07160*, 2023.

GBIF Secretariat. GBIF backbone taxonomy, 2023. URL https://www.gbif.org/dataset/d7dddbf4-2cf0-4f39-9b2a-bb099caae36c.

Gemini Team. Gemini: A family of highly capable multimodal models, 2024. URL https://arxiv.org/abs/2312.11805.

Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio Set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 776–780, 2017. doi: 10.1109/ICASSP.2017.7952261.

Burooj Ghani, Tom Denton, Stefan Kahl, and Holger Klinck. Global birdsong embeddings enable superior transfer learning for bioacoustic classification. *Scientific Reports*, 13(1):22876, 2023.

Yuan Gong, Hongyin Luo, Alexander H Liu, Leonid Karlinsky, and James Glass. Listen, think, and understand. *arXiv preprint arXiv:2305.10790*, 2023.

Masato Hagiwara. AVES: Animal vocalization encoder based on self-supervision. In *ICASSP*, pp. 1–5, 2023. ISBN 978-1-7281-6327-7. doi: 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 *ICASSP*, pp. 1–5, 2023. doi: 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.

Addison Howard, Holger Klinck, Sohier Dane, Stefan Kahl, and Tom Denton. Cornell Birdcall Identification. https://kaggle.com/competitions/birdsong-recognition, 2020. URL https://kaggle.com/competitions/birdsong-recognition. Accessed 2023-06-01.

Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. MobileNets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

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*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Chien-Yu Huang, Ke-Han Lu, Shih-Heng Wang, Chi-Yuan Hsiao, Chun-Yi Kuan, Haibin Wu, Siddhant Arora, Kai-Wei Chang, Jiatong Shi, Yifan Peng, Roshan Sharma, Shinji Watanabe, Bhiksha Ramakrishnan, Shady Shehata, and Hung yi Lee. Dynamic-SUPERB: Towards a dynamic, collaborative, and comprehensive instruction-tuning benchmark for speech, 2024. URL https://arxiv.org/abs/2309.09510.

Rongjie Huang, Mingze Li, Dongchao Yang, Jiatong Shi, Xuankai Chang, Zhenhui Ye, Yuning Wu, Zhiqing Hong, Jiawei Huang, Jinglin Liu, Yi Ren, Zhou Zhao, and Shinji Watanabe. AudioGPT: Understanding and generating speech, music, sound, and talking head, 2023. URL https://arxiv.org/abs/2304.12995.

iNaturalist. iNaturalist. https://www.inaturalist.org/. URL https://www.inaturalist.org/. acccessed 2023-05-01.

M Irfan, Z Jiangbin, S Ali, M Iqbal, Z Masood, and U Hamid. Deepship: An underwater acoustic benchmark dataset and a separable convolution based autoencoder for classification. *Expert Systems with Applications*, 183:115270, 2021.

Stefan Kahl, Connor M. Wood, Maximilian Eibl, and Holger Klinck. BirdNET: A deep learning solution for avian diversity monitoring. *Ecological Informatics*, 61:101236, 2021. ISSN 1574-9541. doi: 10.1016/J.ECOINF.2021.101236.

Stefan Kahl, Russell Charif, and Holger Klinck. A collection of fully-annotated soundscape recordings from the Northeastern United States, September 2022a. URL https://doi.org/10.5281/zenodo.7079380.

Stefan Kahl, Connor M. Wood, Philip Chaon, M. Zachariah Peery, and Holger Klinck. A collection of fully-annotated soundscape recordings from the Western United States, September 2022b. URL https://doi.org/10.5281/zenodo.7050014.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. AudioCaps: Generating captions for audios in the wild. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 119–132, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1011. URL https://aclanthology.org/N19-1011.

Stephanie L King and Vincent M Janik. Bottlenose dolphins can use learned vocal labels to address each other. *Proceedings of the National Academy of Sciences*, 110(32):13216–13221, 2013.

Diederik Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. *Advances in neural information processing systems*, 34:21696–21707, 2021.

Ivan Kiskin, Marianne E. Sinka, Adam D. Cobb, Waqas Rafique, Lawrence Wang, Davide Zilli, Benjamin Gutteridge, Theodoros Marinos, Yunpeng Li, Emmanuel Wilson Kaindoa, Gerard F Killeen, Katherine J. Willis, and S. Roberts. HumBugDB: a large-scale acoustic mosquito dataset. In *Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks*, 2021.

Chun-Yi Kuan, Wei-Ping Huang, and Hung-yi Lee. Understanding sounds, missing the questions: The challenge of object hallucination in large audio-language models. 2024.

Jack LeBien, Ming Zhong, Marconi Campos-Cerqueira, Julian P. Velev, Rahul Dodhia, Juan Lavista Ferres, and T. Mitchell Aide. A pipeline for identification of bird and frog species in tropical soundscape recordings using a convolutional neural network. *Ecological Informatics*, 59:101113, 2020. ISSN 1574-9541. doi: https://doi.org/10.1016/j.ecoinf.2020.101113. URL https://www.sciencedirect.com/science/article/pii/S1574954120300637.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Shansong Liu, Atin Sakkeer Hussain, Chenshuo Sun, and Ying Shan. Music Understanding LLaMA: Advancing text-to-music generation with question answering and captioning. *arXiv preprint arXiv:2308.11276*, 2023.

Siqi Liu, Zhenhai Zhu, Ning Ye, Sergio Guadarrama, and Kevin Murphy. Improved image captioning via policy gradient optimization of SPIDEr. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 873–881, 2017. doi: 10.1109/ICCV.2017.100.

Vincent Lostanlen, Justin Salamon, Mark Cartwright, Brian McFee, Andrew Farnsworth, Steve Kelling, and Juan Pablo Bello. Per-channel energy normalization: Why and how. *IEEE Signal Processing Letters*, 26(1):39–43, 2018.

Xinhao Mei, Chutong Meng, Haohe Liu, Qiuqiang Kong, Tom Ko, Chengqi Zhao, Mark D Plumbley, Yuexian Zou, and Wenwu Wang. WavCaps: A ChatGPT-assisted weakly-labelled audio captioning dataset for audio-language multimodal research. *arXiv*, 2023. doi: 10.48550/arXiv.2303.17395.

Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen. Tut database for acoustic scene classification and sound event detection. In *2016 24th European Signal Processing Conference (EUSIPCO)*, pp. 1128–1132. IEEE, 2016.

Zhongqi Miao, Benjamin Elizalde, Soham Deshmukh, Justin Kitzes, Huaming Wang, Rahul Dodhia, and Juan M. Lavista Ferres. Zero-shot transfer for wildlife bioacoustics detection. *Research Square*, 2023. URL https://doi.org/10.21203/rs.3.rs-3180218/v1.

Zhongqi Miao, Yuanhan Zhang, Zalan Fabian, Andres Hernandez Celis, Sara Beery, Chunyuan Li, Ziwei Liu, Amrita Gupta, Md Nasir, Wanhua Li, Jason Holmberg, Meredith Palmer, Kaitlyn Gaynor, Rahul Dodhia, and Juan Lavista Ferres. New frontiers in AI for biodiversity research and conservation with multimodal language models. *EcoEvoRxiv*, 2024. URL https://ecoevorxiv.org/repository/view/7477/.

Veronica Morfi, Inês Nolasco, Vincent Lostanlen, Shubhr Singh, Ariana Strandburg-Peshkin, Lisa F. Gill, Hanna Pamula, David Benvent, and Dan Stowell. Few-shot bioacoustic event detection: A new task at the DCASE 2021 challenge. In *Detection and Classification of Acoustic Scenes and Events 2021*, 2021.

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

Amanda Navine, Stefan Kahl, Ann Tanimoto-Johnson, Holger Klinck, and Patrick Hart. A collection of fully-annotated soundscape recordings from the Island of Hawai'i, September 2022. URL https://doi.org/10.5281/zenodo.7078499.

Tu Anh Nguyen, Benjamin Muller, Bokai Yu, Marta R. Costa-jussa, Maha Elbayad, Sravya Popuri, Paul-Ambroise Duquenne, Robin Algayres, Ruslan Mavlyutov, Itai Gat, Gabriel Synnaeve, Juan Pino, Benoit Sagot, and Emmanuel Dupoux. SpiRit-LM: Interleaved spoken and written language model, 2024. URL https://arxiv.org/abs/2402.05755.

Guy Oren, Aner Shapira, Reuven Lifshitz, Ehud Vinepinsky, Roni Cohen, Tomer Fried, Guy P. Hadad, and David Omer. Vocal labeling of others by nonhuman primates. *Science*, 385(6712): 996–1003, 2024. doi: 10.1126/science.adp3757. URL https://www.science.org/doi/abs/10.1126/science.adp3757.

Michael A Pardo, Kurt Fristrup, David S Lolchuragi, Joyce H Poole, Petter Granli, Cynthia Moss, Iain Douglas-Hamilton, and George Wittemyer. African elephants address one another with individually specific name-like calls. *Nature Ecology & Evolution*, pp. 1–12, 2024.

Karol J. Piczak. ESC: Dataset for environmental sound classification. In *Proceedings of the 23rd Annual ACM Conference on Multimedia*, pp. 1015–1018. ACM Press, 2015. ISBN 978-1-4503-3459-4. doi: 10.1145/2733373.2806390. URL http://dl.acm.org/citation.cfm?doid=2733373.2806390.

M Poupard, P Best, M Ferrari, P Spong, H Symonds, J-M Prévot, T Soriano, and H Glotin. From massive detections and localisations of orca at orcalab over three years to real-time survey joint to environmental conditions. In *e-Forum Acusticum 2020*, pp. 3235–3237, 2020.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision, 2022. URL https://arxiv.org/abs/2212.04356.

Lukas Rauch, Raphael Schwinger, Moritz Wirth, René Heinrich, Jonas Lange, Stefan Kahl, Bernhard Sick, Sven Tomforde, and Christoph Scholz. Birdset: A multi-task benchmark for classification in avian bioacoustics. *arXiv preprint arXiv:2403.10380*, 2024.

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, ICASSP 2024, Seoul, Republic of Korea, April 14-19, 2024*, pp. 1316–1320. IEEE, 2024. doi: 10.1109/ICASSP48485.2024.10447250. URL https://doi.org/10.1109/ICASSP48485.2024.10447250.

Paul K. Rubenstein, Chulayuth Asawaroengchai, Duc Dung Nguyen, Ankur Bapna, Zalán Borsos, Félix de Chaumont Quitry, Peter Chen, Dalia El Badawy, Wei Han, Eugene Kharitonov, Hannah Muckenhirn, Dirk Padfield, James Qin, Danny Rozenberg, Tara Sainath, Johan Schalkwyk, Matt Sharifi, Michelle Tadmor Ramanovich, Marco Tagliasacchi, Alexandru Tudor, Mihajlo Velimirović, Damien Vincent, Jiahui Yu, Yongqiang Wang, Vicky Zayats, Neil Zeghidour, Yu Zhang, Zhishuai Zhang, Lukas Zilka, and Christian Frank. AudioPaLM: A large language model that can speak and listen. *arXiv preprint arXiv:2306.12925*, 2023.

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://www.science.org/doi/abs/10.1126/science.adg7314.

J Salamon and JP Jacoby, Cand Bello. A dataset and taxonomy for urban sound research. In *Proceedings of the 22nd ACM international conference on Multimedia*, pp. 1041–1044, 2014.

D Santos-Domínguez, S Torres-Guijarro, A Cardenal-López, and A Pena-Gimenez. Shipsear: An underwater vessel noise database. *Applied Acoustics*, 113:64–69, 2016.

Laela Sayigh, Mary Ann Daher, Julie Allen, Helen Gordon, Katherine Joyce, Claire Stuhlmann, and Peter Tyack. The Watkins marine mammal sound database: An online, freely accessible resource. *Proceedings of Meetings on Acoustics*, 27(1):040013, 2016. doi: 10.1121/2.0000358. URL https://asa.scitation.org/doi/abs/10.1121/2.0000358.

Naeha Sharif, Lyndon White, Mohammed Bennamoun, and Syed Afaq Ali Shah. Learning-based composite metrics for improved caption evaluation. In *Proceedings of ACL 2018, student research workshop*, pp. 14–20, 2018.

Petru Soviany, Radu Tudor Ionescu, Paolo Rota, and N. Sebe. Curriculum learning: A survey. *International Journal of Computer Vision*, 130:1526 – 1565, 2021. URL https://api.semanticscholar.org/CorpusID:231709290.

Samuel Stevens, Jiaman Wu, Matthew J Thompson, Elizabeth G Campolongo, Chan Hee Song, David Edward Carlyn, Li Dong, Wasila M Dahdul, Charles Stewart, Tanya Berger-Wolf, Wei-Lun Chao, and Yu Su. BioCLIP: A vision foundation model for the tree of life. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19412–19424, 2024.

Dan Stowell. Computational bioacoustics with deep learning: a review and roadmap. *PeerJ*, 10: e13152, 2022. ISSN 2167-8359. doi: 10.7717/peerj.13152. URL https://europepmc.org/articles/PMC8944344.

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 *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=14rn7HpKVk.

J Thiemann, N Ito, and E Vincent. Diverse environments multichannel acoustic noise database (demand), 2013.

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

Mingqiu Wang, Izhak Shafran, Hagen Soltau, Wei Han, Yuan Cao, Dian Yu, and Laurent El Shafey. Retrieval augmented end-to-end spoken dialog models. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 12056–12060. IEEE, 2024.

Gordon Wichern, Joe Antognini, Michael Flynn, Licheng Richard Zhu, Emmett McQuinn, Dwight Crow, Ethan Manilow, and Jonathan Le Roux. Wham!: Extending speech separation to noisy environments. *arXiv preprint arXiv:1907.01160*, 2019.

Jian Wu, Yashesh Gaur, Zhuo Chen, Long Zhou, Yimeng Zhu, Tianrui Wang, Jinyu Li, Shujie Liu, Bo Ren, Linquan Liu, et al. On decoder-only architecture for speech-to-text and large language model integration. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 1–8. IEEE, 2023a.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP*, pp. 1–5, 2023b. doi: 10.1109/ICASSP49357.2023.10095969.

Xeno-canto. Xeno-canto: Bird sounds from around the world. https://www.xeno-canto.org/. URL https://www.xeno-canto.org/. Accessed 2023-05-15.

Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, Chang Zhou, and Jingren Zhou. AIR-Bench: Benchmarking large audio-language models via generative comprehension. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1979–1998. Association for Computational Linguistics, August 2024. doi: 10.18653/v1/2024.acl-long.109. URL https://aclanthology.org/2024.acl-long.109.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. SpeechGPT: Empowering large language models with intrinsic cross-modal conversational abilities, 2023. URL https://arxiv.org/abs/2305.11000.

Ziqiang Zhang, Sanyuan Chen, Long Zhou, Yu Wu, Shuo Ren, Shujie Liu, Zhuoyuan Yao, Xun Gong, Lirong Dai, Jinyu Li, et al. SpeechLM: Enhanced speech pre-training with unpaired textual data. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

A Appendix

A.1 Xeno-canto held-out species

1. Spice Imperial Pigeon
2. African Pitta
3. New Zealand White-fronted Tern
4. Hume's Treecreeper
5. Brown-rumped Bunting
6. Fiery Minivet
7. Forest Wood Hoopoe
8. Ash-breasted Tit-Tyrant
9. Verreaux's Coua
10. Legge's Hawk-Eagle
11. Red-winged Pytilia
12. Rufous-winged Tanager
13. Forbes-Watson's Swift
14. Blue-chinned Sapphire
15. Moss Frog
16. White-headed Mousebird
17. Tawny-breasted Parrotfinch
18. Ring-tailed Pigeon
19. Pink-backed Pelican
20. Alpine Leaf-Warbler
21. Barred Owlet-nightjar
22. Laurel Pigeon
23. Siberian Blue Robin
24. Yellow-naped Amazon
25. Blue-cheeked Bee-eater
26. Red-knobbed Imperial Pigeon
27. Eurasian Hobby
28. Red-collared Widowbird
29. Northern Red Bishop
30. Shelley's Greenbul
31. Snowy-crowned Robin-Chat
32. Cape Bunting
33. White-crowned Pigeon
34. Sad Flycatcher
35. Asian Dowitcher
36. White-crowned Starling
37. Yellowish White-eye
38. African Silverbill
39. Korean Brown Frog
40. Grey-fronted Honeyeater
41. Red-legged Grasshopper
42. Cook's Robber Frog
43. White-fronted Plover
44. Grey-bellied Squirrel
45. Olive-headed Greenbul
46. Sooty Babbler
47. Large Green Pigeon
48. Red-fronted Rosefinch
49. Bar-breasted Piculet
50. American Black Swift
51. Eurasian Stone-curlew
52. Red-necked Buzzard
53. Streaky-headed Seedeater
54. Rufous Fieldwren
55. Tawny-collared Nightjar
56. Panamanian Flycatcher
57. Black-capped Rufous-Warbler
58. Orange-spotted Bulbul
59. Pere David's Snowfinch
60. Northern Cassowary
61. Yellow-tinted Honeyeater
62. Eastern Tree Frog
63. Frances's Sparrowhawk
64. Sulawesi Swiftlet
65. Gosling's Apalis
66. Eurasian tawny owl
67. Yellow-legged Flyrobin
68. Red-faced Pytilia
69. Double-collared Crescentchest
70. Malagasy Coucal
71. Mountain Bamboo Partridge
72. Zenaida Dove
73. Velvety Black Tyrant
74. Green White-eye
75. Western Rosella
76. Gray Parrot
77. Crested Kingfisher
78. Sunda Owlet
79. Giant Weaver
80. Cape Verde Storm Petrel
81. Rufous-vented Laughingthrush
82. Horned Parakeet
83. Bernier's Teal
84. Sperm Whale
85. Ornate Forest toad
86. Rock Petronia
87. Western Cape Bunting
88. Green Dark Bush-cricket
89. Rufous-cheeked Laughingthrush
90. Scintillant Hummingbird
91. Rufous-webbed Brilliant
92. Handsome Fruiteater
93. Verreaux's Tree Frog
94. Western Black-tailed Rattlesnake
95. Sunda Cuckooshrike
96. Black-crowned Waxbill
97. Whistling Tree Frog
98. Cinderella Waxbill
99. Tawny-backed Fantail
100. Blue-cheeked Flowerpecker
101. Adamawa Turtle Dove
102. Violet-necked Lory
103. Western Orphean Warbler
104. Pacific Robin
105. Black-banded Fruit Dove
106. Black Noddy
107. White-tipped Grasshopper
108. Rusty-necked Piculet
109. Citrine Canary-flycatcher
110. Melancholy Woodpecker
111. La Selle Thrush
112. Cassin's Hawk-Eagle
113. Red-winged Wood Rail
114. Eastern Bristlebird
115. Common Blue-cheeked Bee-eater
116. Grey Cuckooshrike
117. Mottled Duck
118. Bismarck Whistler
119. Black-capped Apalis
120. Indian Skimmer
121. Little Black Cormorant
122. Vaillant's Frog
123. Amazonian Inezia
124. Great Grebe
125. Chestnut-backed Sparrow-Lark
126. Sumba Jungle Flycatcher
127. Tepui Toucanet
128. Elegant Forest Tree Frog
129. Black Guan
130. Pied-winged Swallow
131. Indian Nuthatch
132. McConnell's Spinetail
133. Nepal House Martin
134. Providence Petrel
135. Grey-bellied Shrike-Tyrant
136. Black-necked Grebe
137. Venezuelan Bristle Tyrant
138. Donaldson Smith's Sparrow-Weaver
139. Blyth's Kingfisher
140. Sunset Lorikeet
141. European Golden Plover
142. Biak Monarch
143. Banasura Laughingthrush
144. D'Arnaud's Barbet
145. Tepui Tinamou
146. Lafresnaye's Piculet
147. Fischer's Turaco
148. Christmas White-eye
149. Sooty-capped Hermit
150. Rufous-winged Cisticola
151. Versicolored Barbet
152. Cobb's Wren
153. Black-headed Rufous Warbler
154. Green-throated Mountaingem
155. Knob-billed Fruit Dove
156. Red-eyed Firetail
157. Short-tailed Emerald
158. Sooty Bushtit
159. Bougainville Crow
160. Blue Chaffinch
161. White-winged Scoter
162. Grey-banded Mannikin
163. Giant Antpitta
164. Collared Inca
165. Chilean Skua
166. Rufous-browed Tyrannulet
167. Tanimbar Megapode
168. Thekla Lark
169. Rufous-bellied Euphonia
170. Bannerman's Sunbird
171. Crescent Honeyeater
172. Grey-headed Lovebird
173. Madagascar Snipe
174. Fork-tailed Storm Petrel
175. Armenian Gull
176. Fan-tailed Gerygone
177. Superb Pitta
178. Great White Pelican
179. Huanren Frog
180. Blood-breasted Flowerpecker
181. Margaret's Batis
182. Russet-winged Schiffornis
183. Socotra Cormorant

184. Golden-crowned Emerald
185. Juan Fernandez Petrel
186. Sri Lanka Thrush
187. Golden-winged Sparrow
188. Cream-breasted Fruit Dove
189. Spectacled Tetraka
190. Moluccan Woodcock
191. Yellow-billed Spoonbill
192. Grant's Wood Hoopoe
193. White-fronted Tern
194. Pectoral-patch Cisticola
195. Band-tailed Guan
196. Cameroon Greenbul
197. Eurasian Spoonbill
198. Dusky Babbler
199. Pink Robin
200. Brown Skua
201. Southern Tchagra
202. Great Hornbill
203. Tacarcuna Wood Quail
204. African Wolf
205. Western Cattle Egret
206. Sumatran Woodpecker
207. Eastern Grass Owl
208. Ayacucho Thistletail
209. Philippine Hawk-Eagle
210. Purple-crowned Fairywren
211. Black-faced Babbler
212. Kolombangara Monarch
213. White-browed Treecreeper
214. Emerald Green Tree Frog
215. Cameroon Sunbird
216. Orange-winged Pytilia
217. Tawny Fish Owl
218. Rufous Chatterer
219. White-throated Tapaculo
220. South American Common Toad
221. Cape Streaky-head Seedeater
222. Heuglin's Masked Weaver
223. Dusky White-eye
224. Little Woodpecker
225. Crimson Topaz
226. Glaucous Tanager
227. Ash-throated Casiornis
228. Spotted Wood Owl
229. Atiu Swiftlet
230. Rose-throated Tanager
231. Black-capped Lory
232. Red-breasted Paradise Kingfisher
233. Cinnamon-sided Hummingbird
234. Black Tinamou
235. Striated Wren-Babbler
236. Red-breasted Paradise-Kingfisher
237. Bumpy Rocket Frog
238. Brown Falcon
239. Venezuelan Sylph
240. White-bridled Finch
241. Grey-headed Piprites
242. Western Green Toad
243. South Moluccan Pitta
244. Bornean Black Magpie
245. Western Alpine Mannikin
246. European Herring Gull
247. Cebu Flowerpecker
248. Western Tree Cricket
249. Yellow-knobbed Curassow
250. Flame-throated Sunangel
251. Bare-faced Bulbul
252. Western Grasswren
253. Rufous-vented Chachalaca
254. Pacific Gull
255. Little Sparrowhawk
256. Fine-spotted Woodpecker
257. African Black Swift
258. Pulitzer's Longbill
259. Fast-calling tree cricket
260. bow-winged grasshopper
261. Eirunepe Snouted Tree Frog
262. Caspian Plover
263. Pugnosed Tree Frog
264. Crowned Chat-Tyrant
265. Fire-tailed Sunbird
266. Scaly Babbler
267. Rufous-breasted Warbling Finch
268. Ivory-backed Woodswallow
269. Two-banded Puffbird
270. Buru Golden Bulbul
271. Dusky Gerygone
272. White-breasted Whistler
273. Blackbird
274. Bar-bellied Woodcreeper
275. Socotra Sparrow
276. Grey-bellied Bulbul
277. Cinnamon Tanager
278. Cuban Bullfinch
279. Eye-ringed Flatbill
280. Sooty Antbird
281. Chilean Tinamou
282. China-Muntjak
283. Yellow Rail
284. Luzon Hornbill
285. Everett's White-eye
286. Seram Boobook
287. Bali Myna
288. Green-backed Woodpecker
289. Southern Spotless Crake
290. Choco Tinamou
291. Black-bellied Malkoha
292. Grey-backed Sparrow-Lark
293. Winchell's Kingfisher
294. Maranon Pigeon
295. Violet Wood Hoopoe
296. Grey-hooded Sunbird
297. Common Grasshopper Warbler
298. Tanimbar Starling
299. Southern Variable Pitohui
300. Fairy Tern
301. Carunculated Fruit Dove
302. Erect-crested Penguin
303. California Gull
304. Pallas's Rosefinch
305. Great Gray Owl
306. Kenrick's Starling
307. Brown-winged Parrotbill
308. Green-breasted Bushshrike
309. Green-backed Whistler
310. Fernando Po Batis
311. Chestnut Teal
312. Black Flying Fox
313. Olive-colored White-eye
314. Yellow-headed Amazon
315. Northern Sooty Woodpecker
316. White-lored Antpitta

A.2 Ablation on Unfreezing BEATs

Configuration	watkins	cbi	unseen-cmmn	unseen-sci
BEATs-unfrozen	0.60	0.58	0.08	0.11
BEATs-frozen	0.59	0.35	0.04	0.07

Table 7: Zero-shot classification results with BEATs unfrozen vs. frozen. Both models are trained on stage-1 tasks for 200k steps. We report accuracy on species classification tasks.

A.3 SPEECH+MUSIC ABLATION: FULL RESULTS

Model	esc50	watkins	cbi	humbugdb	dcase	enabirds	hiceas	rfcx	gibbons
base	0.513	0.676	0.702	0.101	0.060	0.257	0.101	0.044	0.010
no-speech-or-music	0.505	0.687	0.705	0.054	0.047	0.259	0.053	0.034	0.010

Table 8: Zero-shot classification and detection results on BEANS-Zero. Base model was trained on all stage-2 training tasks, while no-speech-or-music is an ablation removing both speech and music tasks from training data. Both models were trained for 200k steps. We used accuracy for classification, and F1 for detection tasks.

Model	unseen-cmn	unseen-sci	lifestage	call-type	captioning	zf-indv
base	0.104	0.189	0.661	0.853	0.483	0.379
no-speech-or-music	0.100	0.164	0.700	0.835	0.484	0.243

Table 9: Zero-shot results on new tasks introduced in BEANS-Zero. Base model was trained on all stage-2 training tasks, while no-speech-or-music is an ablation removing both speech and music tasks from training data. Both models were trained for 200k steps. We report accuracy for classification, and SPIDEr (Sharif et al., 2018) for captioning.