Model Merging Improves Zero-Shot Generalization in Bioacoustic Foundation Models

Davide Marincione[1,*] **Donato Crisostomi**[1] **Roberto Dessi**[2]

Emanuele Rodolà[1] **Emanuele Rossi**[1,*]

[1]Department of Computer Science, Sapienza University of Rome [2]Not Diamond, San Francisco, USA

Abstract

Foundation models capable of generalizing across species and tasks represent a promising new frontier in bioacoustics, with NATURELM-AUDIO being one of the most prominent examples. While its domain-specific fine-tuning yields strong performance on bioacoustic benchmarks, we observe that it also introduces trade-offs in instruction-following flexibility. For instance, NATURELM-AUDIO achieves high accuracy when prompted for either the common or scientific name individually, but its accuracy drops significantly when both are requested in a single prompt. We address this by applying a simple model merging strategy that interpolates NATURELM-AUDIO with its base language model, recovering instruction-following capabilities with minimal loss of domain expertise. Finally, we show that the merged model exhibits markedly stronger zero-shot generalization, achieving over a 200% relative improvement and setting a new state-of-the-art in closed-set zero-shot classification of unseen species.

gladia-research-group/model-merging-NatureLM-audio

1 Introduction

Bioacoustics, the study of sound production, transmission, and perception in animals, is a critical tool for understanding biodiversity, monitoring ecosystems, and informing conservation efforts [27, 25, 28]. Recent advances in machine learning (ML) have transformed the field, enabling automated detection, classification, and analysis of acoustic events at unprecedented scales [46].

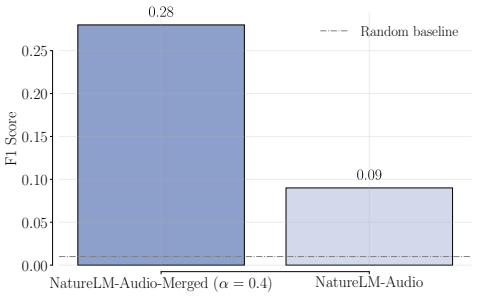


Figure 1: Model merging leads to a *200% relative improvement* over NATURELM-AUDIO in zero-shot classification of unseen species, setting a new state-of-the-art.

Early work in ML for bioacoustics typically relied on *species-specific models*, trained and optimized for a single species and task [1]. However, as in other domains of ML, there is now a shift towards *general-purpose foundation models* that can support a broad range of downstream species and/or tasks with minimal retraining [24, 13, 16, 37, 39, 48]. One of the most prominent examples of this trend is NATURELM-AUDIO [40], the first bioacoustics audio–language model, designed for zero-shot generalization to unseen tasks via text-based prompting.

In this paper, we examine the capabilities of NATURELM-AUDIO as a *general* foundation model for bioacoustics. Despite its strong performance on tasks and prompts closely matching its training distribution, we find that its intense domain-specific fine-tuning has led to a severe reduction in

*Equal technical contribution

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: AI for Non-Human Animal Communication.

Common Name Prompt

```
What is the common name for the focal species
in the audio?
```

Scientific Name Prompt

```
What is the scientific name for the focal
species in the audio?
```

Combined Prompt

```
Identify the focal species in the audio and
provide its scientific name, followed by a
colon and its common name.
```

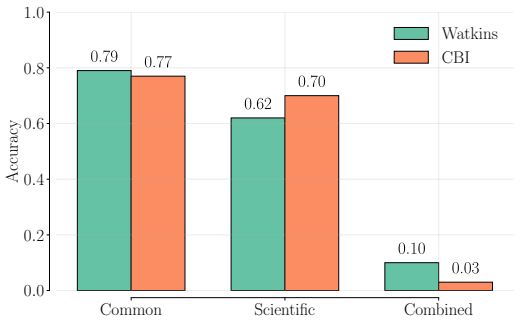


Figure 2: NATURELM-AUDIO classification accuracy for different prompts on WATKINS and CBI.

id aecd4535-ebef-438d-811b-1ffb7be5c22e

```
GT: Odobenus rosmarus: Walrus
Common: Walrus
Scientific: Odobenus rosmarus
Combined: Odobenus rosmarus - male courtship ...
```

id 4cbdd583-23c6-408d-aea5-3719dc6d9654

```
GT: Lagenodelphis hosei: Frasers Dolphin
Common: Fraser's Dolphin
Scientific: Lagenodelphis hosei
Combined: Lagenodelphis hosei
```

id 185c974a-d905-475c-8fc0-0cbab1e383b9

```
GT: Eubalaena australis: Southern Right Whale
Common: Fin- Finback Whale
Scientific: Balaenoptera physalus
Combined: Balaenoptera physalus: 52 Hz Pulses
```

id 4186fc55-0ac2-4f44-9026-009f239fcf96

```
GT: Stenella clymene: Clymene Dolphin
Common: Clymene Dolphin
Scientific: Stenella clymene
Combined: Stenella clymene: Clymene Dolphin
```

Figure 3: Example model predictions for the common name, scientific name, and combined-name prompts, compared to ground truth. Correct predictions in green, incorrect in red.

the *instruction-following capabilities* of its base model (LLAMA-3.1-8B-INSTRUCT), a trade-off commonly observed in other specialized models [56] and limiting its ability to generalize in zero-shot settings. We show that *model merging* with the base model can help restore these capabilities, achieving a balance between domain-specific knowledge and general instruction-following. Finally, we show that this approach sets a new *state-of-the-art in zero-shot classification of unseen species*, achieving over a 200% relative improvement compared to NATURELM-AUDIO (figure 1).

2 Problem Analysis

NATURELM-AUDIO is a LoRA [19] fine-tuning of LLAMA-3.1-8B-INSTRUCT [14] on ∼2M steps of audio–text pairs, predominantly bioacoustics but also music and human sounds. While the original evaluation shows that the model follows training-like instructions well, such as predicting either the common or the scientific name of the focal species in an audio in isolation, we find that requesting both in a single prompt often leads to a large drop in accuracy. Figure 2 shows the prompts and corresponding accuracies on WATKINS and CBI, two species-classification datasets from the BEANS-ZERO benchmark [17] covering marine mammals and birds, respectively. On both datasets, the model performs better on common names than on scientific names, yet achieves high accuracy (60–80%) in both cases. However, when prompted for both names jointly, accuracy falls to 6–12%.

The examples in figure 3 illustrate typical failure modes. In the top left, the model outputs correct common and scientific names individually, but under the combined prompt it drifts into behavioural description ("male courtship behavior"), possibly reflecting its exposure to captioning-style data during training [40]. In the bottom left, it misidentifies the species in all cases, yet common and scientific predictions remain mutually consistent; in the combined case it again appends unrelated context ("52 Hz pulses"). In the top right, correct individual predictions degrade to only the scientific name under the combined prompt. In the bottom right, all three prompts succeed.

We additionally experiment with the ZF-INDIV dataset originally used in [40] to evaluate zero-shot task generalization (see appendix B.3) and observe a similar pattern: NATURELM-AUDIO shows

reduced robustness to even mild prompt variations. This behaviour is consistent with the effects of extensive domain-specific fine-tuning observed in other specialized LLMs, where overfitting to training prompt formats can narrow instruction-following flexibility and limit generalization [56].

3 Method

Section 2 shows that NATURELM-AUDIO has lost its instruction following capabilities in favor of task-specific ones acquired during fine-tuning. We recover these ones through model merging.

Model Merging Model merging aims to ensemble different models without incurring in additional inference or storage costs [2, 12, 51, 5]. While the non-linear nature of neural networks prevents from taking the weighted average of the models in general [50], this aggregation is well behaved when the two models exhibit linear mode connectivity [10], *i.e.* can be connected via a linear path over which the loss does not significantly increase. In this case, the merged model $\mathbf{\Theta}^{(\text{merge})}$ can be obtained from the endpoint models $\mathbf{\Theta}^{(1)}, \mathbf{\Theta}^{(2)}$ simply as $\mathbf{\Theta}^{(\text{merge})} = (1-\alpha)\mathbf{\Theta}^{(1)} + \alpha\mathbf{\Theta}^{(2)}$, where $\alpha \in [0, 1]$ is a scaling parameter controlling the contribution of each model. Consistent with previous findings [50, 33, 10], we observe that linear interpolation remains effective along the fine-tuning trajectory, suggesting that linear mode connectivity holds when (part of) the optimization path is shared.

Merging NATURELM-AUDIO with its base model We merge LLAMA-3.1-8B-INSTRUCT with its fine-tuning NATURELM-AUDIO to combine the instruction following abilities of the former and the task-specific performance of the latter. In particular, being NATURELM-AUDIO a LoRa [19] fine-tuning, linearly interpolating between the base and the fine-tuned is equivalent to changing the multiplicative factor α in LoRa: Given the weight matrix $\mathbf{W}_{\text{base}}$ of the base model for some layer, LoRA updates it as $\mathbf{W}_{\text{ft}} = \mathbf{W}_{\text{base}} + \mathbf{AB}$, where $\mathbf{A}$ and $\mathbf{B}$ are two low-rank learnable matrices; thus

$$(1-\alpha)\,\mathbf{W}_{\text{base}} + \alpha\,\mathbf{W}_{\text{ft}} = (1-\alpha)\,\mathbf{W}_{\text{base}} + \alpha\,(\mathbf{W}_{\text{base}} + \mathbf{AB}) \tag{1}$$

$$= \mathbf{W}_{\text{base}} - \cancel{\alpha\,\mathbf{W}_{\text{base}}} + \cancel{\alpha\,\mathbf{W}_{\text{base}}} + \alpha\,\mathbf{AB} = \mathbf{W}_{\text{base}} + \alpha\,\mathbf{AB}. \tag{2}$$

This shows that we can interpolate between the base and the fine-tuned model simply by varying α.

4 Results

Combined Instruction-Following Task We evaluate the merged model over a range of interpolation coefficients α, using the three prompt variants in figure 2. The y-axis in figure 4 reports the accuracy on the *combined* prompt, while the x-axis shows the mean accuracy on the *training-like* prompts (common name and scientific name individually). For the combined prompt, intermediate interpolation values substantially outperform both extremes. Specifically, rescaling from $\alpha = 1$ (NATURELM-AUDIO) to $\alpha \approx 0.7$ increases combined-task accuracy from $6\% \rightarrow 45\%$ on Watkins and $12\% \rightarrow 63\%$ on CBI, reflecting a restoration of instruction-following capabilities degraded in the fine-tuned model. However, setting α too low ($\alpha < 0.5$) sharply reduces accuracy on combined prompts due to a loss of domain-specific audio knowledge from the fine-tuning stage.

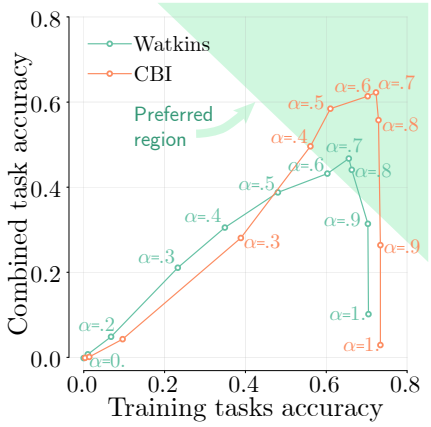


Figure 4: Accuracy on the combined prompt (y-axis) from figure 2 versus the mean accuracy on the individual common- and scientific-name prompts (x-axis) when varying α.

The observed behaviour highlights α as a controllable *capability trade-off parameter*. At $\alpha = 1$, the model fully retains its domain adaptation but suffers in compositional instruction following. At $\alpha = 0$, it maximizes general instruction-following behaviour inherited from the base model, but discards most bioacoustic specialization. The monotonic decline in x-axis accuracy with decreasing α further confirms that domain-task performance and instruction-following ability are in tension.

Classification of Unseen Species We next evaluate the merged model on the task of *zero-shot classification of previously unseen species*, using the UNSEEN-FAMILY-CMN split of the BEANS-

ZERO benchmark [40]. This split, comprising 425 samples from 40 classes, enforces that no species from the same taxonomic family appear in both training and test sets. The task is to predict each focal species' *common name* without any fine-tuning. Unlike the open-vocabulary formulation of BEANS-ZERO, we adopt a *closed-set* setup where all possible 40 target labels are provided in the prompt, allowing us to assess adherence to label constraints. This setting also mirrors realistic field conditions where the set of possible species is geographically bounded.

Quantitative results (figure 1) show that the merged model achieves a *200% relative improvement* over NATURELM-AUDIO (F1 = 0.28 vs. 0.09), demonstrating markedly stronger generalization to unseen taxa and setting a new state-of-the-art. To better understand this gain, we analyze a simplified binary version of the task where the model has to predict only among the two most popular classes (125 samples, figure 5). For $\alpha \in [0.6, 1.0]$, the model rarely confuses valid classes but frequently produces out-of-set predictions, signaling strong discrimination but weak prompt adherence. As α decreases, these invalid outputs vanish, reaching an optimal trade-off around $\alpha = 0.4$, beyond which in-set confusions and abstentions slightly increase. The results suggest that the NATURELM-AUDIO audio encoder already provides robust species-level representations, and that its main bottleneck lies in the language model's capacity to follow prompts. Model merging effectively mitigates this limitation and greatly improves *closed-set zero-shot classification* by enhancing adherence to the prompt constraints.

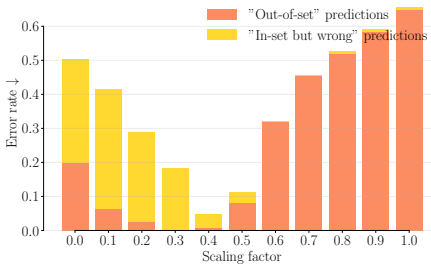


Figure 5: Error breakdown (lower is better) on the binary subset of UNSEEN-FAMILY-CMN as a function of the merging coefficient α.

5 Related Work

Catastrophic forgetting in multi-modal fine-tuning Catastrophic forgetting is a well-known challenge when fine-tuning large language models [41]. A common training-time mitigation is to freeze the LLM and update only the projection layer that maps visual or audio features into the text embedding space, often with fewer fine-tuning steps [56] or using PEFT methods [57, 35]. In contrast, post-training approaches aim to restore forgotten skills in already fine-tuned models [59, 36].

Model merging Model merging provides an efficient alternative to ensembling, producing a single model combining multiple models' capabilities without increasing inference cost. Early work, inspired by linear mode connectivity [10, 12, 31, 9], focused on aligning independently trained models, often by solving a neuron permutation problem [2, 23, 5, 42, 43, 15]. Closer to our work, Wortsman et al. [50] produce robust fine-tuned models by linearly interpolating them with their base model, while Ilharco et al. [21] use interpolations to improve specific tasks without waiving others.

6 Conclusions

We investigated the instruction-following limitations of NATURELM-AUDIO and found that even small changes in prompt structure can significantly degrade performance, reducing its utility as a general-purpose model. To address this, we applied a lightweight model-merging strategy that interpolates the fine-tuned NATURELM-AUDIO with its base model. Intermediate interpolation weights restore much of the lost instruction-following capability while preserving most domain-specific accuracy. This recovery further improves zero-shot classification of unseen species, setting a new state-of-the-art. In our experiments, $\alpha \in [0.4, 0.6]$ provided a strong balance between instruction following and domain expertise, though the optimal value remains task- and dataset-dependent.

Limitations and Future Work Our current evaluation of zero-shot closed-set classification is limited in scope and we plan to extend it to multiple datasets in the future. Convex weight interpolation may not be optimal, and we intend to explore alternative strategies for restoring instruction-following abilities, including more advanced model-merging methods [3, 30, 11, 26, 44]) and activation-steering techniques [4, 45].

Acknowledgments

We thank Emmanuel Chemla for his helpful feedback. This work is partly supported by the MUR FIS2 grant n. FIS-2023-00942 "NEXUS" (cup B53C25001030001), and by Sapienza University of Rome via the Seed of ERC grant "MINT.AI" (cup B83C25001040001).

References

[1] T.M̃itchell Aide, C. Corrada-Bravo, M. Campos-Cerqueira, C. Milan, G. Vega, and R. Alvarez. Real-time bioacoustics monitoring and automated species identification. *PeerJ*, 1:e103, 2013. doi: 10.7717/peerj.103. URL https://doi.org/10.7717/peerj.103.

[2] Samuel K. Ainsworth, Jonathan Hayase, and Siddhartha S. Srinivasa. Git re-basin: Merging models modulo permutation symmetries. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*, 2023.

[3] Takuya Akiba, Makoto Shing, Yujin Tang, Qi Sun, and David Ha. Evolutionary optimization of model merging recipes. *Nature Machine Intelligence*, 2025. ISSN 2522-5839.

[4] Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. *Advances in Neural Information Processing Systems*, 37:136037–136083, 2024.

[5] Donato Crisostomi, Marco Fumero, Daniele Baieri, Florian Bernard, and Emanuele Rodolà. C^2M^3: Cycle-consistent multi-model merging. In *Advances in Neural Information Processing Systems*, volume 37, 2025.

[6] Nico Daheim, Thomas Möllenhoff, Edoardo Ponti, Iryna Gurevych, and Mohammad Emtiyaz Khan. Model merging by uncertainty-based gradient matching. In *The Twelfth International Conference on Learning Representations*.

[7] MohammadReza Davari and Eugene Belilovsky. Model breadcrumbs: Scaling multi-task model merging with sparse masks. In *European Conference on Computer Vision*. Springer, 2025.

[8] Julie E Elie and Frédéric E Theunissen. The vocal repertoire of the domesticated zebra finch: a data-driven approach to decipher the information-bearing acoustic features of communication signals. *Anim Cogn*, 19: 285–315, 2015.

[9] Rahim Entezari, Hanie Sedghi, Olga Saukh, and Behnam Neyshabur. The role of permutation invariance in linear mode connectivity of neural networks. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*, 2022.

[10] Jonathan Frankle, Gintare Karolina Dziugaite, Daniel Roy, and Michael Carbin. Linear mode connectivity and the lottery ticket hypothesis. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, Proceedings of Machine Learning Research, 2020.

[11] Antonio Andrea Gargiulo, Donato Crisostomi, Maria Sofia Bucarelli, Simone Scardapane, Fabrizio Silvestri, and Emanuele Rodolà. Task singular vectors: Reducing task interference in model merging. In *Proc. CVPR*, 2025.

[12] Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry P. Vetrov, and Andrew Gordon Wilson. Loss surfaces, mode connectivity, and fast ensembling of dnns. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, 2018.

[13] 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.

[14] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[15] Fidel A. Guerrero-Peña, Heitor Rapela Medeiros, Thomas Dubail, Masih Aminbeidokhti, Eric Granger, and Marco Pedersoli. Re-basin via implicit sinkhorn differentiation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, 2023.

[16] Masato Hagiwara. Aves: Animal vocalization encoder based on self-supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[17] Masato Hagiwara, Benjamin Hoffman, Jen-Yu Liu, Maddie Cusimano, Felix Effenberger, and Katie Zacarian. Beans: The benchmark of animal sounds. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[18] Stefan Horoi, Albert Manuel Orozco Camacho, Eugene Belilovsky, and Guy Wolf. Harmony in Diversity: Merging Neural Networks with Canonical Correlation Analysis. 2024.

[19] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

[20] Chenyu Huang, Peng Ye, Tao Chen, Tong He, Xiangyu Yue, and Wanli Ouyang. Emr-merging: Tuning-free high-performance model merging, 2024.

[21] Gabriel Ilharco, Mitchell Wortsman, Samir Yitzhak Gadre, Shuran Song, Hannaneh Hajishirzi, Simon Kornblith, Ali Farhadi, and Ludwig Schmidt. Patching open-vocabulary models by interpolating weights. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.

[22] Gabriel Ilharco, Marco Túlio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*, 2023.

[23] Keller Jordan, Hanie Sedghi, Olga Saukh, Rahim Entezari, and Behnam Neyshabur. REPAIR: renormalizing permuted activations for interpolation repair. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*, 2023.

[24] Stefan Kahl, Connor M Wood, Maximilian Eibl, and Holger Klinck. Birdnet: A deep learning solution for avian diversity monitoring. *Ecological Informatics*, 61:101236, 2021.

[25] Paola Laiolo. The emerging significance of bioacoustics in animal species conservation. *Biological conservation*, 143(7):1635–1645, 2010.

[26] Daniel Marczak, Simone Magistri, Sebastian Cygert, Bartłomiej Twardowski, Andrew D Bagdanov, and Joost van de Weijer. No task left behind: Isotropic model merging with common and task-specific subspaces. In *Forty-second International Conference on Machine Learning*.

[27] Peter R Marler and Hans Slabbekoorn. *Nature's music: the science of birdsong*. Elsevier, 2004.

[28] 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.

[29] Michael Matena and Colin Raffel. Merging models with fisher-weighted averaging. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.

[30] Tommaso Mencattini, Robert Adrian Minut, Donato Crisostomi, Andrea Santilli, and Emanuele Rodolà. Merge^3: Efficient evolutionary merging on consumer-grade gpus. In *Forty-second International Conference on Machine Learning*, 2025.

[31] Seyed-Iman Mirzadeh, Mehrdad Farajtabar, Dilan Görür, Razvan Pascanu, and Hassan Ghasemzadeh. Linear mode connectivity in multitask and continual learning. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*, 2021.

[32] Aviv Navon, Aviv Shamsian, Ethan Fetaya, Gal Chechik, Nadav Dym, and Haggai Maron. Equivariant deep weight space alignment, 2023.

[33] Behnam Neyshabur, Hanie Sedghi, and Chiyuan Zhang. What is being transferred in transfer learning? *Advances in neural information processing systems*, 33:512–523, 2020.

[34] Guillermo Ortiz-Jiménez, Alessandro Favero, and Pascal Frossard. Task arithmetic in the tangent space: Improved editing of pre-trained models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

[35] Ashwinee Panda, Berivan Isik, Xiangyu Qi, Sanmi Koyejo, Tsachy Weissman, and Prateek Mittal. Lottery ticket adaptation: Mitigating destructive interference in llms. *ArXiv preprint*, 2024.

[36] Abhishek Panigrahi, Nikunj Saunshi, Haoyu Zhao, and Sanjeev Arora. Task-specific skill localization in fine-tuned language models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 27011–27033. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/panigrahi23a.html.

[37] Lukas Rauch, René Heinrich, Ilyass Moummad, Alexis Joly, Bernhard Sick, and Christoph Scholz. Can masked autoencoders also listen to birds?, 2025. URL https://arxiv.org/abs/2504.12880.

[38] Peter Robicheaux, Matvei Popov, Anish Madan, Isaac Robinson, Joseph Nelson, Deva Ramanan, and Neehar Peri. Roboflow100-vl: A multi-domain object detection benchmark for vision-language models. *Advances in Neural Information Processing Systems*, 2025.

[39] David Robinson, Adelaide Robinson, and Lily Akrapongpisak. Transferable models for bioacoustics with human language supervision, 2023. URL https://arxiv.org/abs/2308.04978.

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

[41] Haizhou Shi, Zihao Xu, Hengyi Wang, Weiyi Qin, Wenyuan Wang, Yibin Wang, and Hao Wang. Continual learning of large language models: A comprehensive survey. *ACM Computing Surveys*, 2024. URL https://api.semanticscholar.org/CorpusId:269362836.

[42] Sidak Pal Singh and Martin Jaggi. Model fusion via optimal transport. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[43] George Stoica, Daniel Bolya, Jakob Brandt Bjorner, Pratik Ramesh, Taylor Hearn, and Judy Hoffman. Zipit! merging models from different tasks without training. In *The Twelfth International Conference on Learning Representations*, .

[44] George Stoica, Pratik Ramesh, Boglarka Ecsedi, Leshem Choshen, and Judy Hoffman. Model merging with svd to tie the knots. In *The Thirteenth International Conference on Learning Representations*, .

[45] Alessandro Stolfo, Vidhisha Balachandran, Safoora Yousefi, Eric Horvitz, and Besmira Nushi. Improving instruction-following in language models through activation steering. In *The Thirteenth International Conference on Learning Representations*.

[46] Dan Stowell. Computational bioacoustics with deep learning: a review and roadmap. *PeerJ*, 10:e13152, 2022.

[47] Anke Tang, Li Shen, Yong Luo, Yibing Zhan, Han Hu, Bo Du, Yixin Chen, and Dacheng Tao. Parameter-efficient multi-task model fusion with partial linearization. In *The Twelfth International Conference on Learning Representations*.

[48] Bart van Merriënboer, Vincent Dumoulin, Jenny Hamer, Lauren Harrell, Andrea Burns, and Tom Denton. Perch 2.0: The bittern lesson for bioacoustics, 2025. URL https://arxiv.org/abs/2508.04665.

[49] Ke Wang, Nikolaos Dimitriadis, Guillermo Ortiz-Jimenez, François Fleuret, and Pascal Frossard. Localizing task information for improved model merging and compression. In *Proceedings of the 41st International Conference on Machine Learning*, Proceedings of Machine Learning Research, 2024.

[50] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo-Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. Robust fine-tuning of zero-shot models, 2021.

[51] Mitchell Wortsman, Gabriel Ilharco, Samir Yitzhak Gadre, Rebecca Roelofs, Raphael Gontijo Lopes, Ari S. Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, Proceedings of Machine Learning Research, 2022.

[52] Prateek Yadav, Derek Tam, Leshem Choshen, Colin A. Raffel, and Mohit Bansal. Ties-merging: Resolving interference when merging models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

[53] Enneng Yang, Li Shen, Zhenyi Wang, Guibing Guo, Xiaojun Chen, Xingwei Wang, and Dacheng Tao. Representation surgery for multi-task model merging. In *Proceedings of the 41st International Conference on Machine Learning*, Proceedings of Machine Learning Research, 2024.

[54] Enneng Yang, Zhenyi Wang, Li Shen, Shiwei Liu, Guibing Guo, Xingwei Wang, and Dacheng Tao. Adamerging: Adaptive model merging for multi-task learning. In *The Twelfth International Conference on Learning Representations*, 2024.

[55] Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. Language models are super mario: Absorbing abilities from homologous models as a free lunch. In *Proceedings of the 41st International Conference on Machine Learning*, Proceedings of Machine Learning Research, 2024.

[56] Yuexiang Zhai, Shengbang Tong, Xiao Li, Mu Cai, Qing Qu, Yong Jae Lee, and Yi Ma. Investigating the catastrophic forgetting in multimodal large language model fine-tuning. In *Conference on Parsimony and Learning (Proceedings Track)*, 2023. URL https://openreview.net/forum?id=g7rMSiNtmA.

[57] Weixiang Zhao, Shilong Wang, Yulin Hu, Yanyan Zhao, Bing Qin, Xuanyu Zhang, Qing Yang, Dongliang Xu, and Wanxiang Che. Sapt: A shared attention framework for parameter-efficient continual learning of large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *ACL (1)*, pages 11641–11661. Association for Computational Linguistics, 2024. ISBN 979-8-89176-094-3. URL http://dblp.uni-trier.de/db/conf/acl/acl2024-1.html#ZhaoWHZQZYXC24.

[58] Luca Zhou, Daniele Solombrino, Donato Crisostomi, Maria Sofia Bucarelli, Fabrizio Silvestri, and Emanuele Rodolà. Atm: Improving model merging by alternating tuning and merging, 2024. URL https://arxiv.org/abs/2411.03055.

[59] Didi Zhu, Zhongyi Sun, Zexi Li, Tao Shen, Ke Yan, Shouhong Ding, Chao Wu, and Kun Kuang. Model tailor: mitigating catastrophic forgetting in multi-modal large language models. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

A Extended related work

Foundation Models in Bioacoustics Recent advances have introduced large-scale bioacoustic foundation models designed for cross-species and cross-task generalization. NATURELM-AUDIO [40] integrates a self-supervised audio encoder with a LLaMA-based language decoder. BIOLINGUAL [39] adapts CLAP-style audio–text contrastive learning to bioacoustics, while audio-only models such as BIRDMAE [37], AVES [16] and PERCH 2.0 [48] pretrain large models on extensive birdsong or multi-taxa datasets to produce broadly transferable acoustic features. Although these models surpass species-specific baselines, they remain susceptible to domain shifts and catastrophic forgetting, which limits their robustness in real-world deployments.

Mode connectivity and model merging Mode connectivity investigates the weight configurations that define local minima. Frankle et al. [10] examines the linear mode connectivity of models trained for only a few epochs from the same initialization, linking this phenomenon to the lottery ticket hypothesis. Relaxing the shared-initialization requirement, Entezari et al. [9] argues that, after resolving neuron permutations [32, 18], all trained models may reside in a single connected basin. Model merging pursues a different goal: combining multiple models into one that inherits their capabilities without the cost and complexity of ensembling. In this direction, Singh and Jaggi [42] introduced an optimal-transport–based weight-matching method, while `Git Re-Basin` [2] proposes optimizing a linear assignment problem (LAP) for each layer. More recently, `REPAIR` [23] shows that substantial gains in the performance of interpolated models can come from renormalizing activations, while C^2M^3 [5] proposes matching and merging many models jointly through cycle-consistent permutations. When the models to merge are fine-tuned from a shared backbone, task-vector-based methods are most effective [22, 52, 55, 29, 50, 7, 49, 58, 11, 34, 30, 20, 6, 44, 53, 47]. These involve taking the parameter-level difference between the fine-tuned model and its pretrained base, termed a task vector. Improvements can be obtained by optimizing task-vector combinations [54], mitigating sign disagreement [52], randomly dropping updates [55], or applying evolutionary methods [3, 30]. Finally, techniques employing layer-wise task vectors [44, 11, 26] obtain state-of-the-art results by leveraging layer-level structures through SVD of the parameter differences.

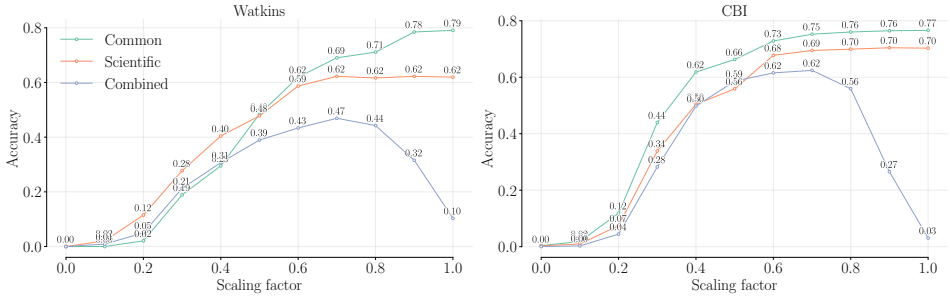


Figure 6: Accuracy on the common name, scientific name, and combined prompts from figure 2, as a function of the rescaling parameter α.

B Additional Experiments

B.1 Combined Instruction Following Task

Figure 2 shows that the accuracy of the original NATURELM-AUDIO model on the "combined prompt" is drastically lower than its "common" and "scientific" counterparts, and figure 4 shows how the accuracy of the "combined prompt" varies as a function of the accuracy over the other two prompts. Here, Figure 6 provides a more granular visualization, explicitly plotting how the accuracy of each prompt evolves with changes in the rescaling parameter α, thereby highlighting the trade-off between instruction-following recovery and domain-specific retention.

B.2 Experimental Details

Correct prediction estimation Following [40], we evaluate classification accuracy by first extracting the model's free-form output and then computing the Levenshtein distance between this output and each possible target class name (in this case, the two species' common names). The class with the smallest distance is selected, and the prediction is considered correct if this distance is less than a threshold t (set to $t = 5$ in our experiments).

For small α, we often observe that the model tends to output a consistent prefix sentence before presenting the actual answer. When prompted for the common/scientific name (on any experiment) the model responds with "`The common/scientific name for the focal species in the audio is {species_name}`", and specifically, when a scientific name is requested, the model sometimes also follows with "`, also known as {common_name_of_species}`". Due to the formulaic nature of these answers, we decided to check for such cases and extract the model's actual prediction from them, before computing the above-mentioned distance computation (as otherwise, they would be flagged as incorrect).

Classification of Unseen Species To mitigate position bias, we randomly permute the order of few-shot examples for each evaluation sample. This randomization is applied independently for every sample, thereby minimizing spurious correlations between class position and model predictions. As previously noted, the two species used in the experiment, Spotted Elachura (*Elachura formosa*) and Dall's Porpoise (*Phocoenoides dalli*), were selected for being the most frequent classes in the UNSEEN-FAMILY-CMN dataset, with 73 and 53 samples respectively.

B.3 Zero-Shot Generalization Task

In Robinson et al. [40], zero-shot generalization was evaluated on the ZF-INDIV dataset [8], part of the BEANS-ZERO benchmark [17], as it was excluded from the model's training set. ZF-INDIV tests the ability to infer the number of Zebra Finch (*Taeniopygia guttata*) individuals in an audio recording, making it a particularly relevant downstream task.

With the original prompt from Robinson et al. [40] (figure 9), NATURELM-AUDIO achieves 0.66 accuracy (random baseline: 0.5), indicating partial generalization to this unseen task. However,

Closed-Set Classification Prompt

```
What is the common name for the focal species in the audio?  Output exactly one of:  {species_list}
```

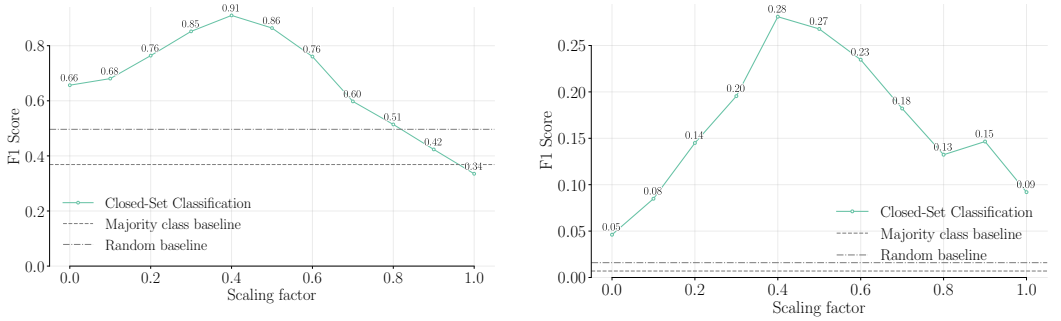


Figure 7: F1 classification score on UNSEEN-CMN-FAMILY when varying α. In the left plot, only the two most popular classes are kept for the task, while on the right plot all classes are used. Note the different scales of the plots.

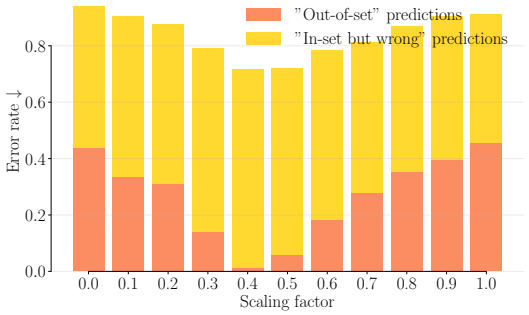


Figure 8: Error breakdown (lower is better) on UNSEEN-FAMILY-CMN as a function of the merging coefficient α.

reversing the class name order or removing explicit class labels (see figure 9 for the exact prompts), reduces accuracy to 0.52, effectively random performance.

These results suggest that the higher-than-random performance reported in Robinson et al. [40] is sensitive to prompt formulation. While the original result remains valid for the tested prompt, our findings indicate that the apparent zero-shot generalization may partly stem from prompt-specific biases rather than task understanding.

B.3.1 Few-Shot In-Context Learning

Given the merged model's demonstrated ability to follow prompts, we further investigated whether it could also benefit from few-shot in-context examples, a setting particularly relevant to bioacoustics, where practitioners often have access to only a handful of labeled recordings for new species, habitats, or acoustic conditions. Specifically, we evaluated a binary version of the task by directly injecting the audio tokens obtained from the BEATs + Q-Former component of NATURELM-AUDIO into the prompt. We tested prompts containing no examples ($k = 0$, identical to the Closed-Set Classification setup in figure 7), one example per class ($k = 1$), and five examples per class ($k = 5$). The prompts and results are shown in figure 10. Across all α values, in-context examples did not yield consistent gains. Instead, performance often degraded or fluctuated unpredictably, with higher k leading to noisier behavior. This trend mirrors recent findings showing that multimodal large language models frequently fail to benefit from few-shot multimodal prompting, and in some cases (e.g., Gemini 2.5 Pro) such context can even harm detection accuracy [38]. The effect is likely exacerbated by the fact that neither NATURELM-AUDIO nor the base LLAMA-3.1-8B-INSTRUCT model were trained on multi-audio prompts, making setups like figure 10 substantially out-of-distribution, even for interpolated α values.

Original Prompt

```
Is there only one bird in the audio, or more?
Reply with 'One' or 'More'.
```

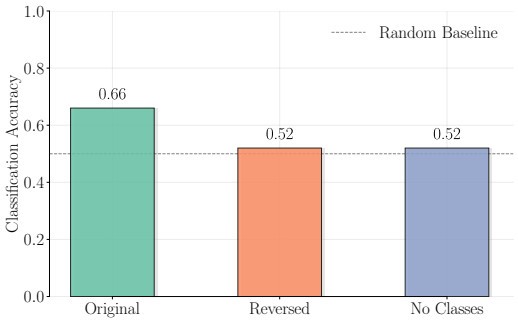
Reversed Classes Prompt

```
Is there more than one bird in the audio, or
just one?  Reply with 'More' or 'One'.
```

No Classes Prompt

```
How many birds are there in the audio?
```

(a)



(b)

Figure 9: **Classification accuracy for different prompt types on ZF-INDIV.** (a) Exact wording of the three evaluated prompts. (b) Accuracy of NATURELM-AUDIO on ZF-INDIV. Accuracy is above random for the original prompt from Robinson et al. [40], but drops to near-random when the prompt is slightly reworded.

In-Context Learning Prompt (One Shot)

```
Identify the common name for the focal species
in the audio.  Output exactly one of:  Dall's
Porpoise, Spotted Elachura

Audio:  <Audio><SampleAudioHere></Audio>
Label:  Dall's Porpoise

Audio:  <Audio><SampleAudioHere></Audio>
Label:  Spotted Elachura

Audio:  <Audio><TestAudioHere></Audio>
Label:
```

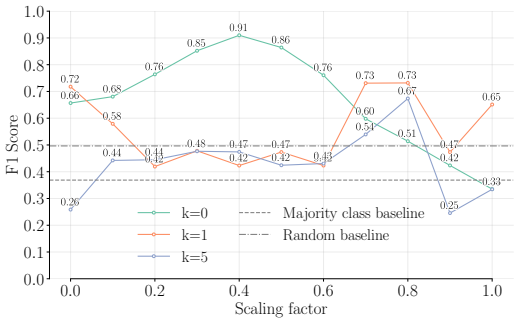


Figure 10: Example prompt and performance for few-shot in-context classification on UNSEEN-FAMILY-CMN when varying the merging coefficient α. The prompt (left) illustrates how few-shot examples are formatted using audio tokens, while the plot (right) reports the resulting F1-scores for $k \in 0, 1, 5$. Adding in-context examples does not consistently improve performance and can lead to noisier behavior across α values.

B.3.2 Compute Resources

All experiments were conducted on one NVIDIA A100 GPU (40 GB), using 8 CPU cores and 32 GB of RAM, requiring 300 GB of disk space, primarily for storing the BEANS-ZERO benchmark. However, since only a fraction of the benchmark was used, the actual memory footprint could be substantially smaller.