

Model Merging Improves Zero-Shot Generalization in Bioacoustic Foundation Models

D. Marincione*, D. Crisostomi, R. Dessì, E. Rodolà, **E. Rossi***

ESP Community Deep Dive, NeurIPS 2025 Workshop

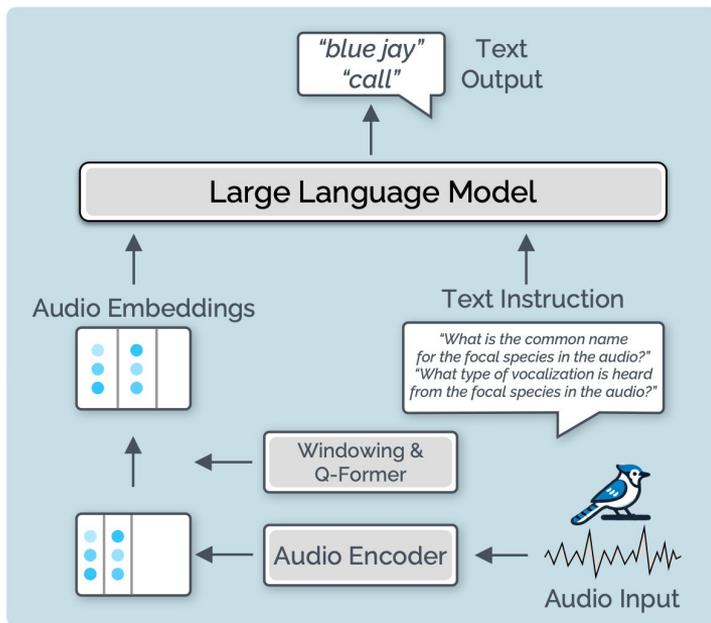
Background

NatureLM: the first bioacoustics audio-language model

| Input | Model Prediction |
|---|---|
|  <p>▶ 0:00 / 0:12 </p> <p>Prompt: What is the common name for the focal species in the audio?</p> | <i>Humpback Whale</i> |
|  <p>▶ 0:00 / 0:07 </p> <p>Prompt: What is the life stage of the focal species in the audio?</p> | <i>juvenile</i> |
|  <p>▶ 0:00 / 0:10 </p> <p>Prompt: Caption the audio, using the common name for any animal species.</p> | <i>Call of a new zealand bellbird with background sounds from new zealand falcon.</i> |

NatureLM: technical details

- Audio encoder
- LLM: Llama-3.1-8B-instruct (fine-tuned with LoRa)



| Task | Audio Input | Text Instruction | Text Output |
|----------------|-------------|---|---|
| Classification | | Which of these is the focal species in the audio? Bachman's Sparrow, Grey Shrikethrush, Unicolored Jay | Grey Shrikethrush |
| Detection | | Which of these, if any, are present in the audio recording? Long-billed Wren, Eurasian Wren, None | Eurasian Wren |
| Captioning | | Caption the audio, using the common name for any animal species | The sound of a Swamp Sparrow trilling twice with a brief gap in between. |

Figure 2: Examples of training instances

Problem Analysis

NatureLM is not robust to prompt variations

- Even small deviations from training prompts lead to large drops in accuracy

Common Name Prompt

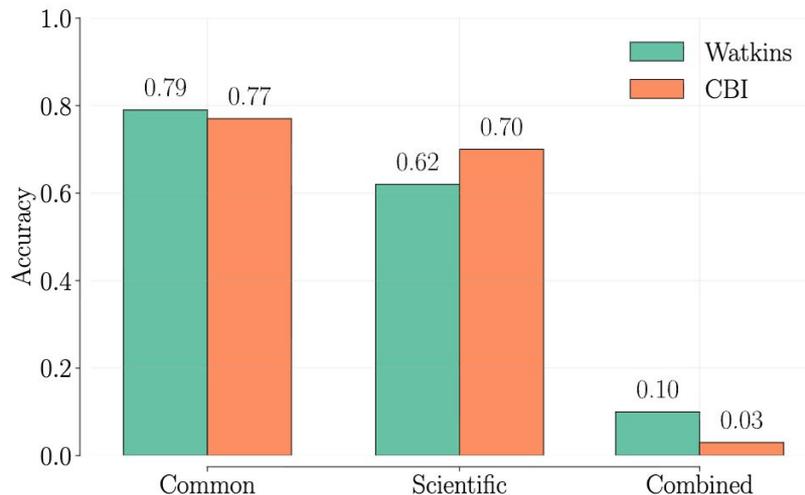
What is the common name for the focal species in the audio?

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.



NatureLM is not able to follow instructions

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

GT: *Stenella clymene*: Clymene Dolphin

Common: **Clymene Dolphin**

Scientific: ***Stenella clymene***

Combined: ***Stenella clymene*: Clymene Dolphin**

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 aece4535-ebef-438d-811b-1ffb7be5c22e

GT: *Odobenus rosmarus*: Walrus

Common: **Walrus**

Scientific: ***Odobenus rosmarus***

Combined: ***Odobenus rosmarus* - male courtship ...**

- NatureLM-Audio has lost its instruction following capabilities in favor of task-specific ones acquired during fine-tuning

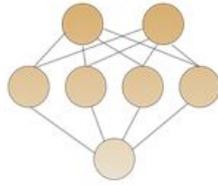
Method and Results

What do we need?

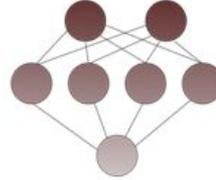
- Improving the instruction following capabilities, while retaining the bioacoustics expertise
- Avoiding the costly process of retraining the model

Model Merging

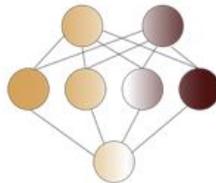
Researcher: solving narrow problems



Navigator: navigating complex systems



University Professor



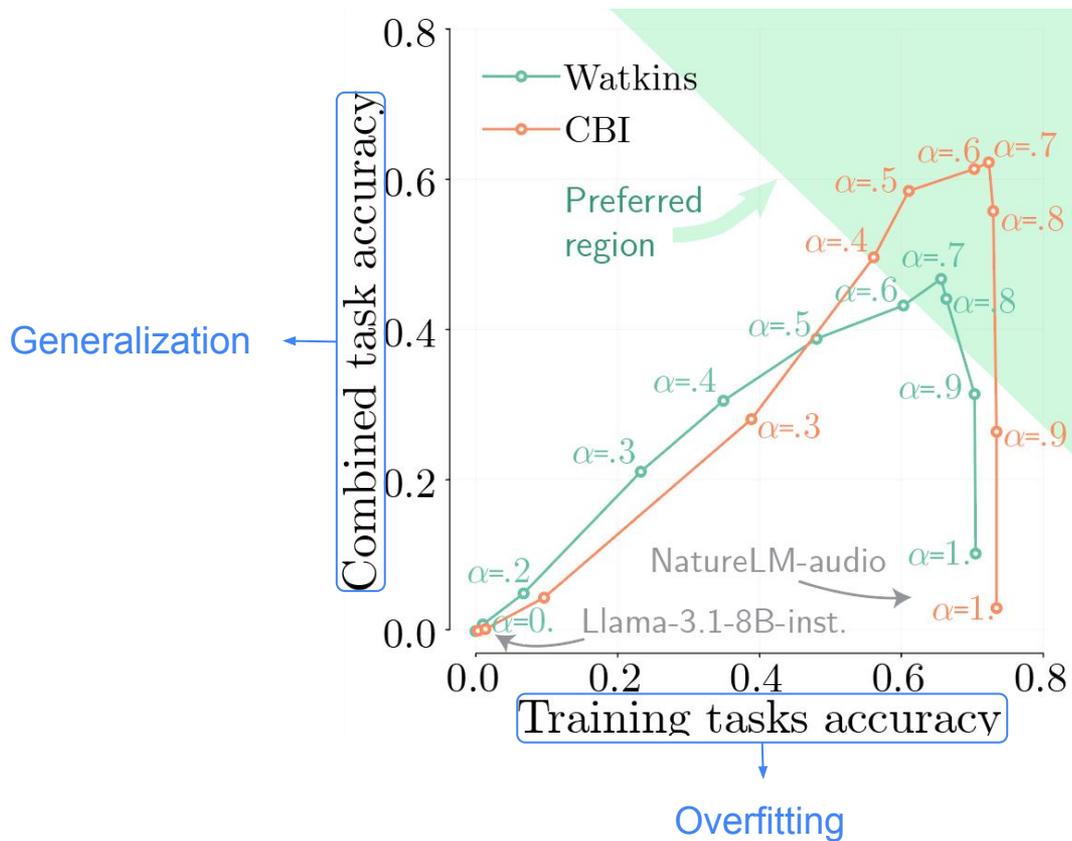
Model Merging with the base model

- We merge the fine-tuned LM with its base model through linear interpolation

$$\begin{aligned}(1 - \alpha) \mathbf{W}_{\text{base}} + \alpha \mathbf{W}_{\text{ft}} &= (1 - \alpha) \mathbf{W}_{\text{base}} + \alpha (\mathbf{W}_{\text{base}} + \mathbf{AB}) \\ &= \mathbf{W}_{\text{base}} - \cancel{\alpha \mathbf{W}_{\text{base}}} + \cancel{\alpha \mathbf{W}_{\text{base}}} + \alpha \mathbf{AB} = \mathbf{W}_{\text{base}} + \alpha \mathbf{AB}.\end{aligned}$$

LoRa update

Model Merging restores the lost capabilities

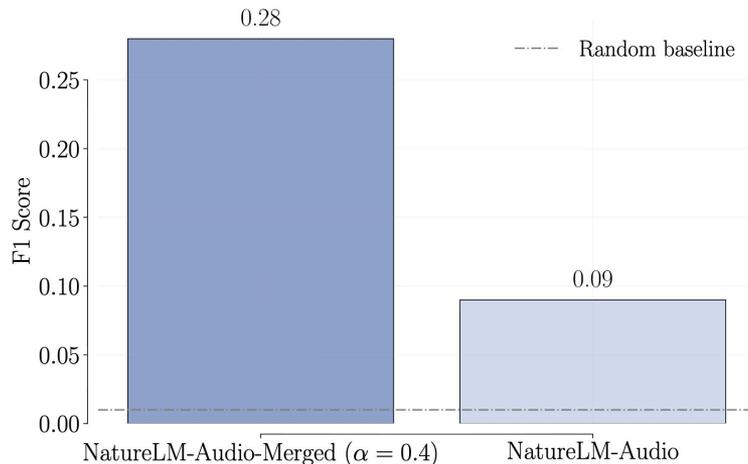


New SOTA on Unseen Species Classification

- Zero-shot closed-set classification of species never seen during training

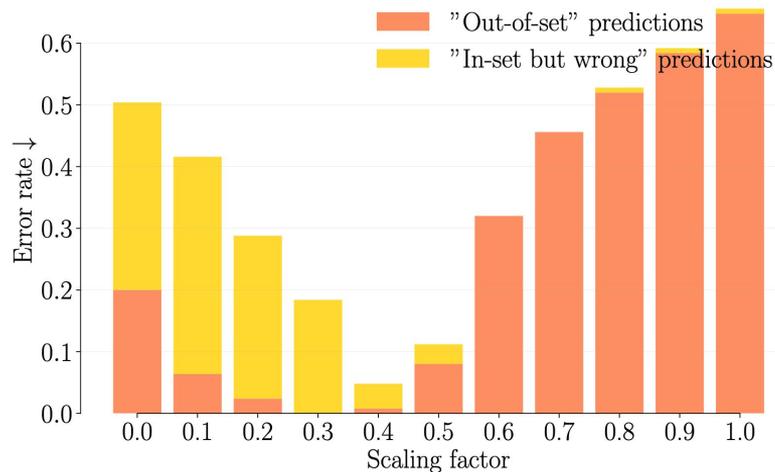
Closed-Set Classification Prompt

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



What is actually going on? Simplified Binary Setup

- NatureLM ($\alpha=1$) often predicts labels outside the provided set
 - causes worse-than-random performance



- Reducing α mitigates this effect, with $\alpha=0.4$ offering the best trade-off

Already Merged into the Main Nature-LM Repo

- Original authors (ESP) found it super useful and it has already been integrated

The screenshot shows a GitHub pull request page for the repository 'earthspecies / NatureLM-audio'. The pull request title is 'Implement Model Merging for Improved Instruction-Following and Zero-Shot Generalization #14'. It is marked as 'Merged' and was merged by GaganNarula into the 'earthspecies:main' branch from the 'emalgorith:feat-model-merging' branch. The pull request includes 4 commits, 0 checks, and 3 files changed, with a net change of +25 lines and 0 deletions. A comment from 'emalgorith' provides a summary and implementation details.

Summary

This PR implements model merging as described in [Model Merging Improves Zero-Shot Generalization in Bioacoustic Foundation Models](#). The implementation allows users to interpolate between NatureLM and its base language model at inference time by scaling LoRA adapter weights, recovering instruction-following capabilities and improving zero-shot generalization.

Implementation

This PR adds a `merging_alpha` parameter to the `GenerateConfig` class that controls the interpolation strength:

- `merging_alpha = 1.0` (default): Uses the full fine-tuned NatureLM model (no merging)
- `merging_alpha < 1.0`: Interpolates toward the base language model (e.g., `0.4` means 40% NatureLM, 60% base model)

Reviewers

- GaganNarula ✓

Assignees

No one assigned

Labels

None yet

Projects

None yet

Milestone

No milestone

Thank you!

