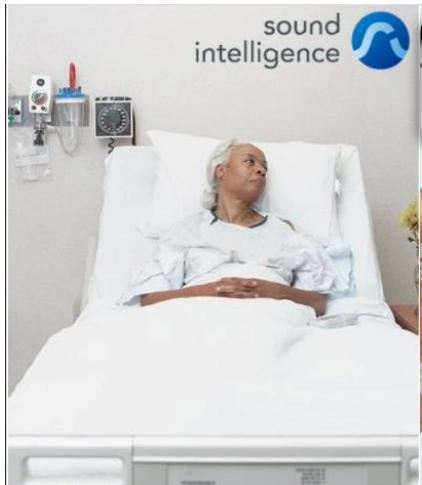


# Learning Audio Foundation Models for Reasoning

Soham Deshmukh  
Carnegie Mellon University

Committee: Bhiksha Raj, Rita Singh, Shinji Watanabe, Dan Ellis, Benjamin Elizalde

# Audio Understanding has multiple applications



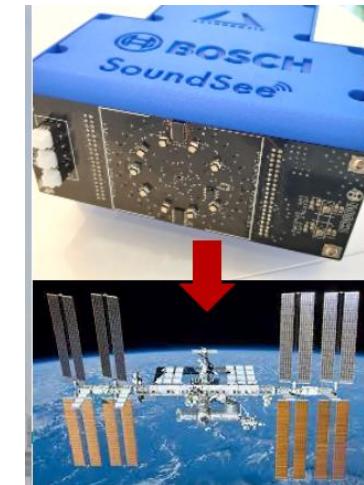
Healthcare



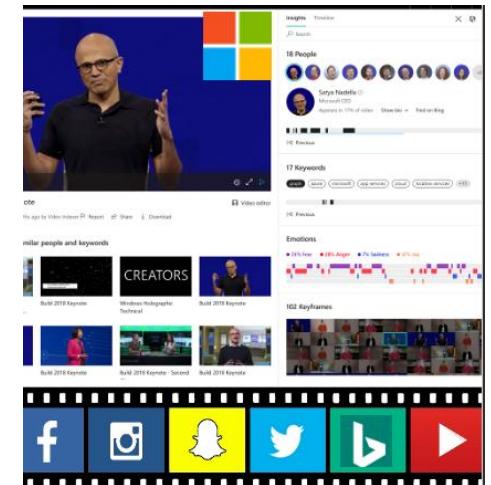
Safety



Noise Monitoring



Predictive Maintenance

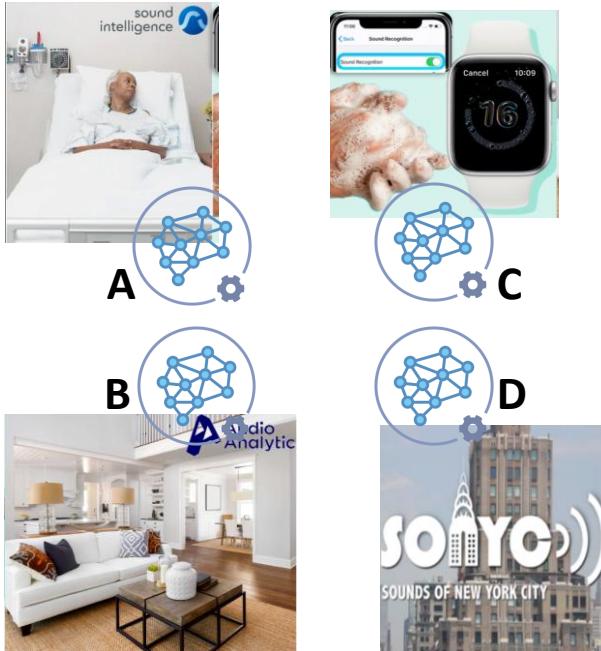


Content Retrieval

# ML models built for each task and domain



# Task-specific to Foundation Models



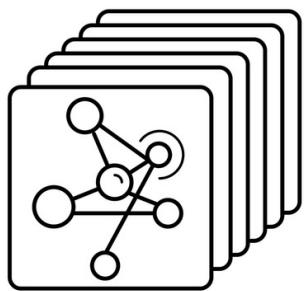
Towards foundation  
model paradigm →



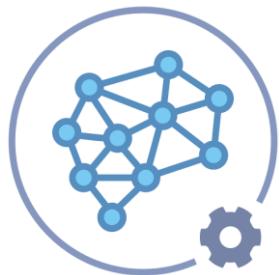
Computer Audition: From Task-Specific Machine Learning to Foundation Models,  
<https://arxiv.org/abs/2407.15672>

# Audio Foundation Model (AFM)

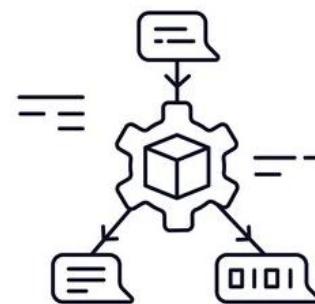
Multi-purpose ML model pre-trained on extensive audio datasets using self-supervised learning (SSL)



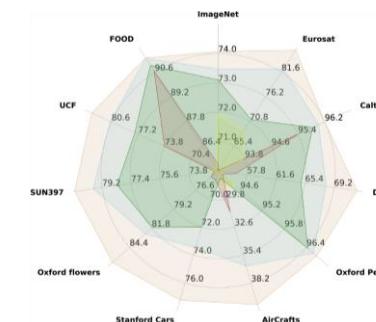
Pretraining on  
large datasets



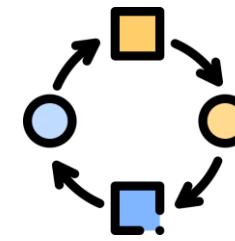
Self-Supervised  
Learning



Multi-modal  
capabilities



Generalization and  
versatility

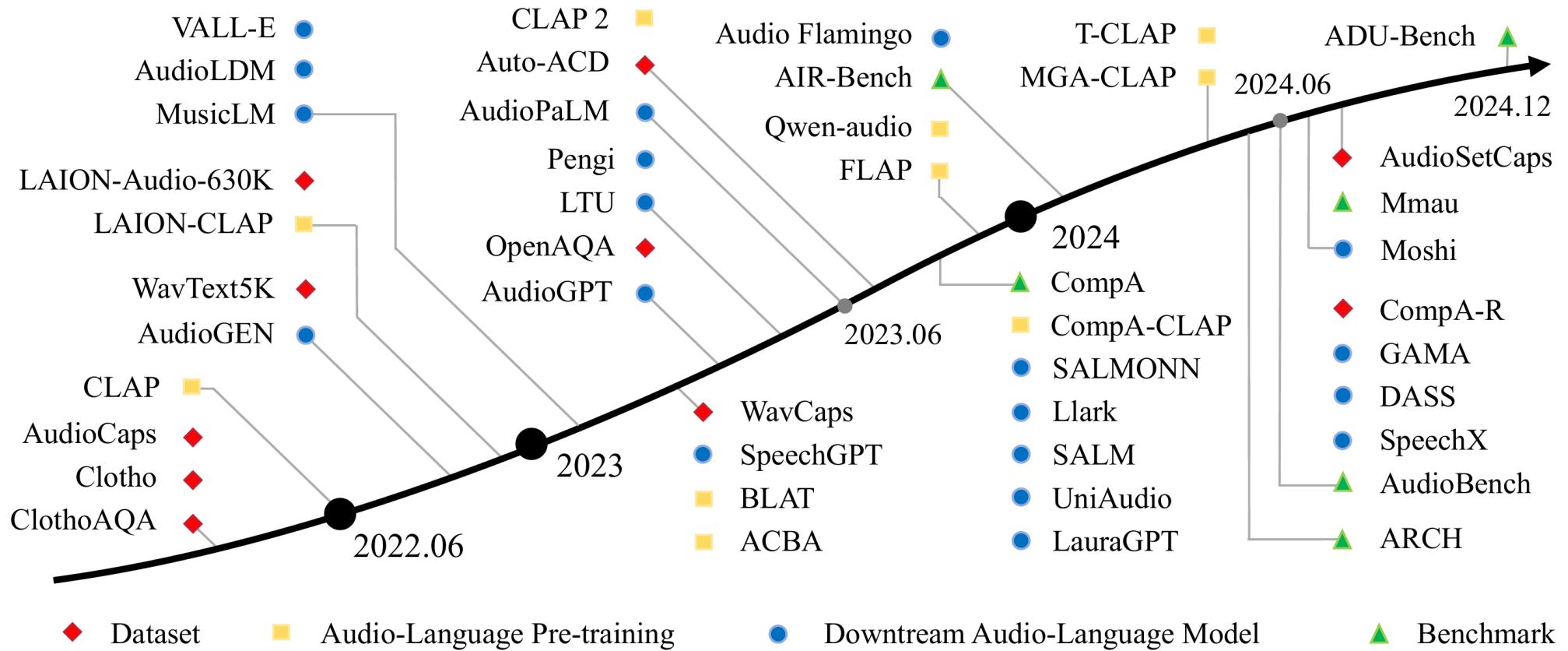


Adaptability



Emergent  
abilities

# AFM research is steadily growing



This thesis investigates two central questions

# This thesis investigates two central questions

1. What are effective learning or pretraining methods to enable the scaling of audio models?

# This thesis investigates two central questions

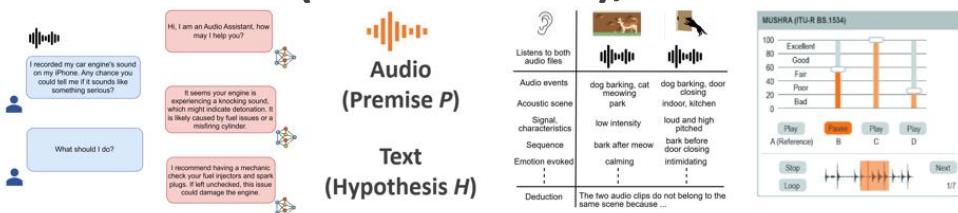
1. What are effective learning or pretraining methods to enable the scaling of audio models?
2. Once pretrained, how can the new abilities be analyzed, evaluated, and subsequently improved?

# Thesis contributions

- Effective learning and pretraining methods  
**CLAP (ICASSP 23, 24), Pengi (NeurIPS 23), Mellow (under review)**



- Methods or benchmarks to analyze new abilities  
**Audio Entail (AAAI 25 Oral), Audio Difference (ICLR 25 Spotlight), PAM (INTERSPEECH 24)**



- Efficient adaptation methods to enable customization  
**Test-time (ICASSP 24), Few-shot (INTERSPEECH 25), Missing modality (ICASSP 25 Oral)**



# Talk outline

- Effective learning and pretraining methods  
**CLAP (ICASSP 23, 24), Pengi (NeurIPS 23), Mellow (under review)**



- Methods or benchmarks to analyze new abilities  
**Audio Entail (AAAI 25 Oral), Audio Difference (ICLR 25 Spotlight), PAM (INTERSPEECH 24)**



- Efficient adaptation methods to enable customization  
**Test-time (ICASSP 24), Few-shot (INTERSPEECH 25), Missing modality (ICASSP 25)**



# Talk outline

- **Effective learning and pretraining methods**

CLAP (ICASSP 23, 24), Pengi (NeurIPS 23), Mellow (under review)



- Methods or benchmarks to analyze new abilities

Audio Entail (AAAI 25 Oral), Audio Difference (ICLR 25 Spotlight), PAM (INTERSPEECH 24)



- Efficient adaptation methods to enable customization

Test-time (ICASSP 24), Few-shot (INTERSPEECH 25), Missing modality (ICASSP 25)



# A typical process for building audio classifier

## Collect data

Dog barking



## Define classifier

Sound Event  
Classifier

## Training

Helicopter

River

Dog barking

Wind

Laughing

# A typical process for building audio classifier

## Collect data

Dog barking



## Define classifier

Sound Event  
Classifier

## Training

Helicopter

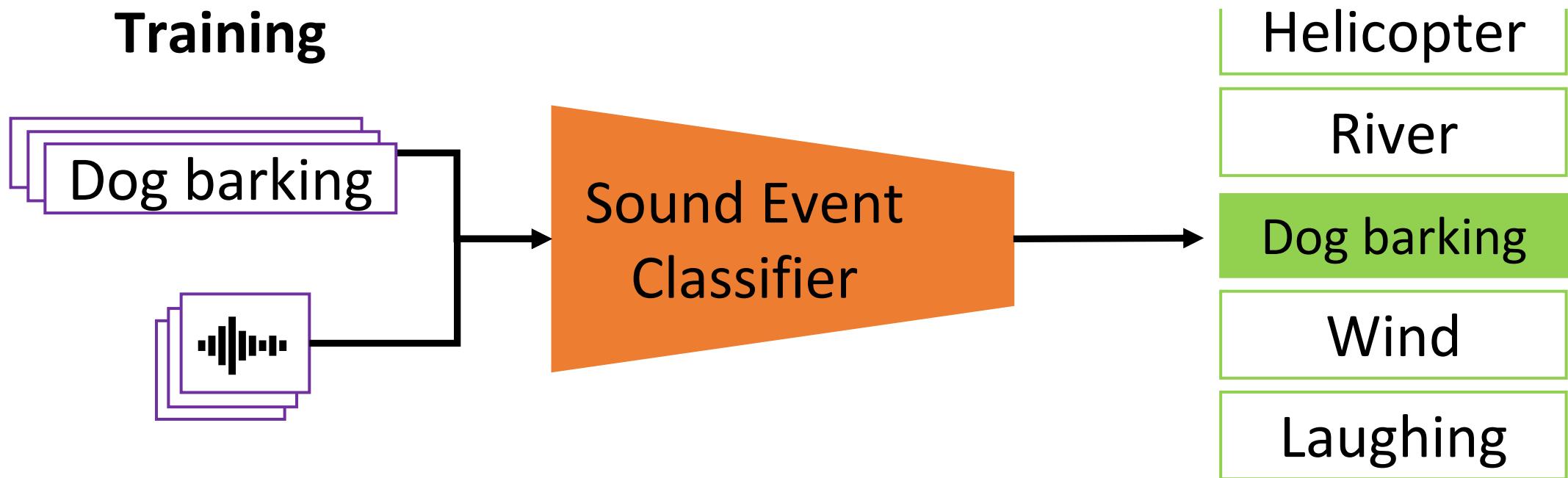
River

Dog barking

Wind

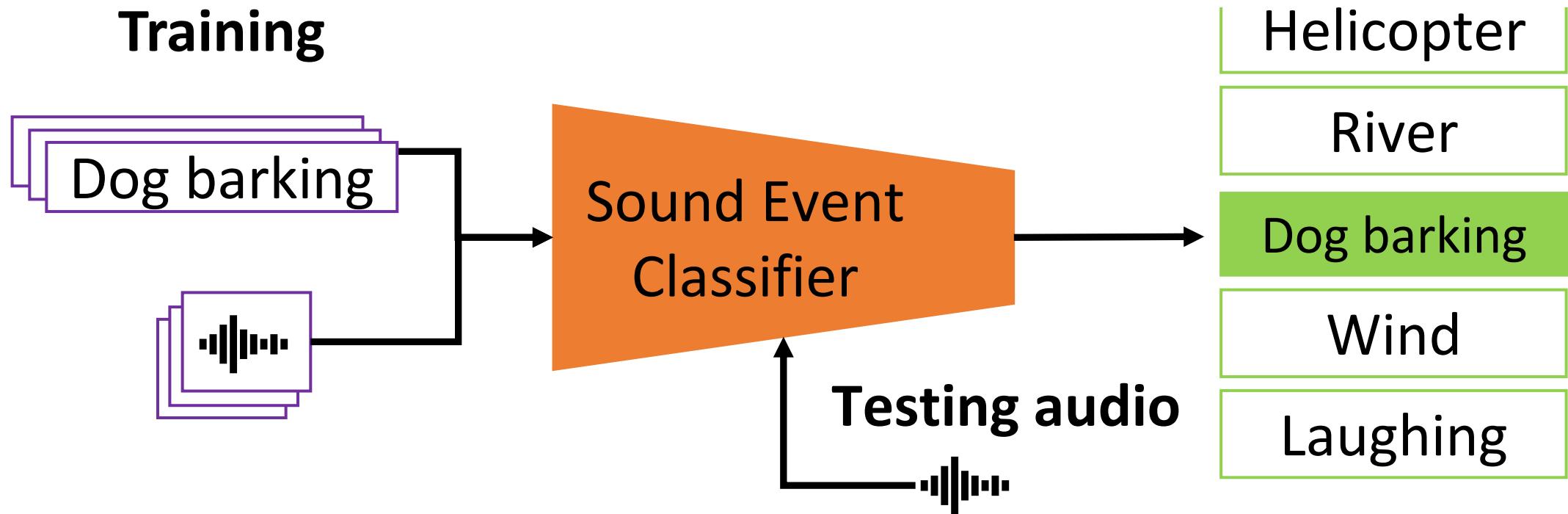
Laughing

# Supervised model training



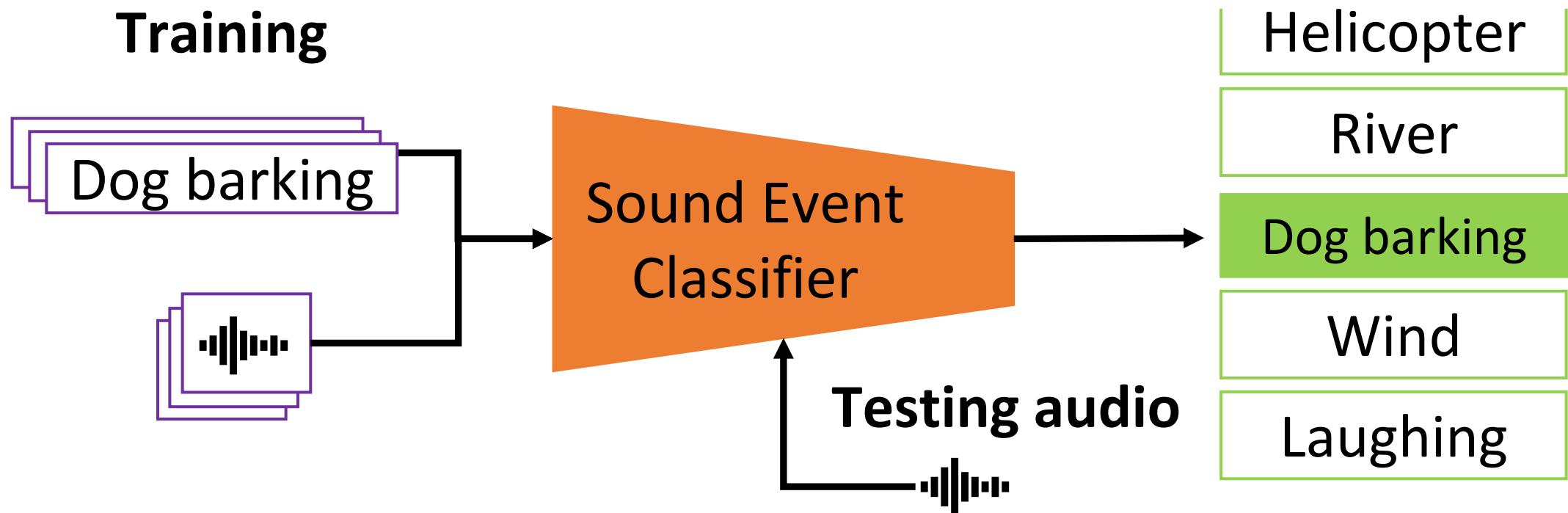
Collecting annotated data and training model using  
supervised learning

# Predicts predefined classes



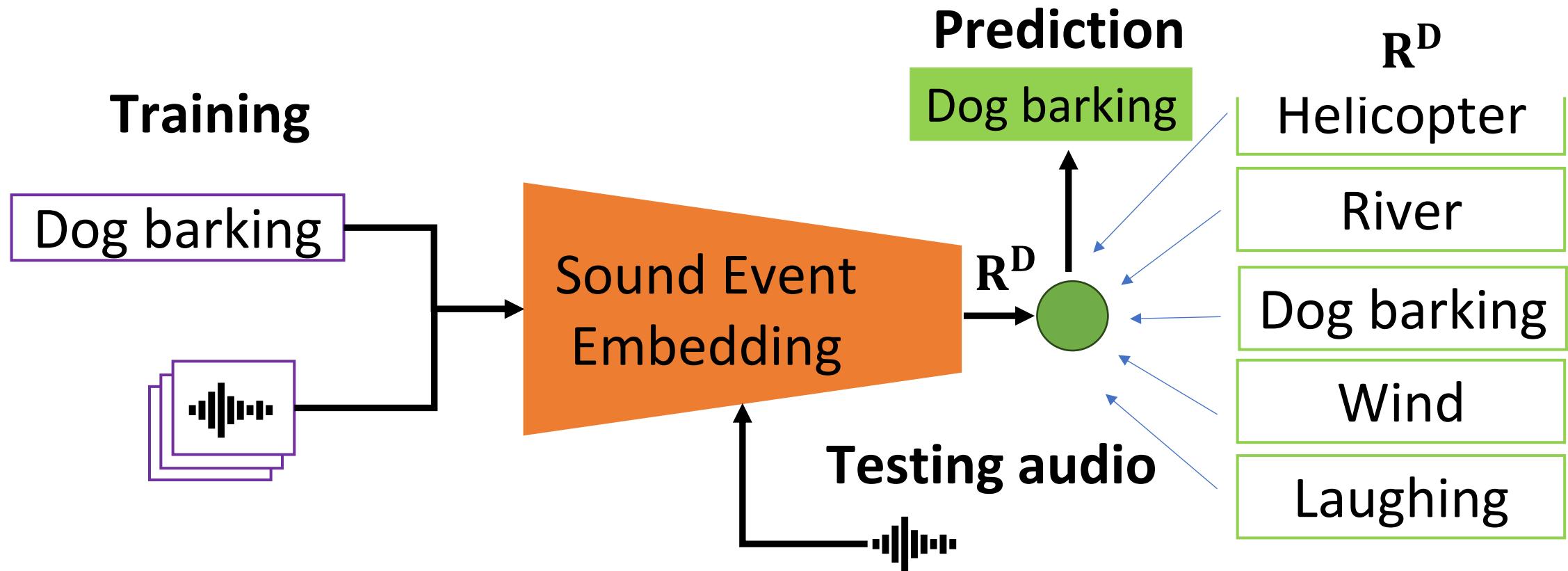
During inference, the model predicts 1 out of N classes

# Predicts predefined classes



How to overcome predefined classes

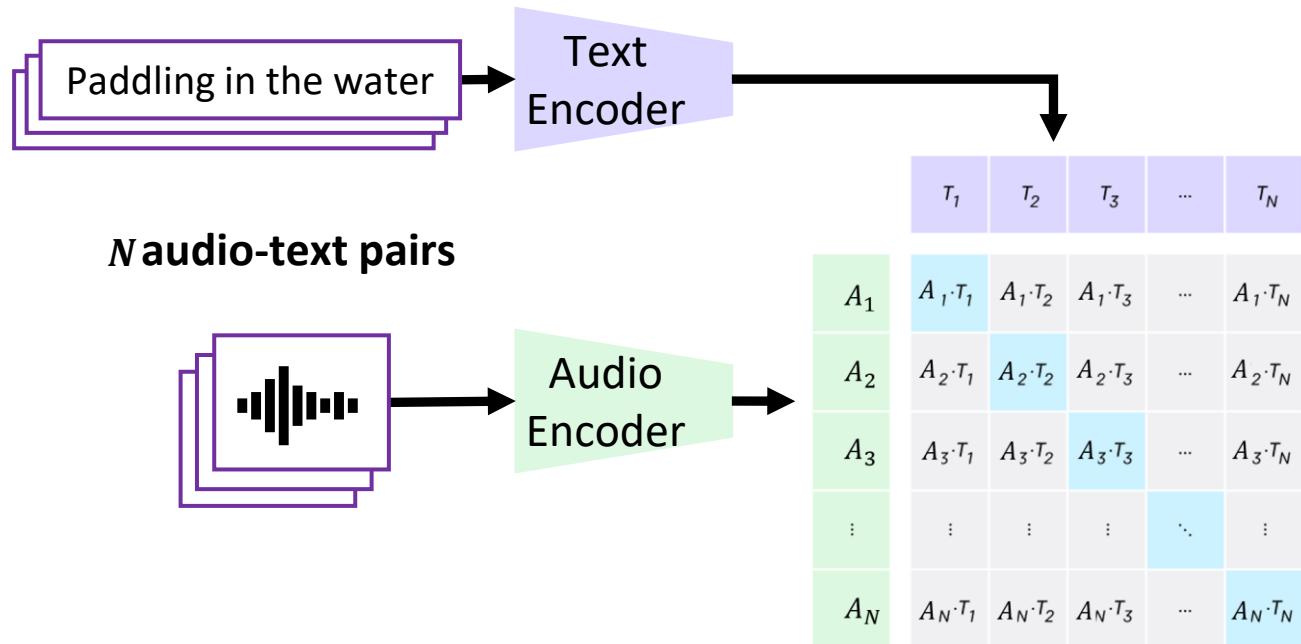
# Overcoming predefined classes



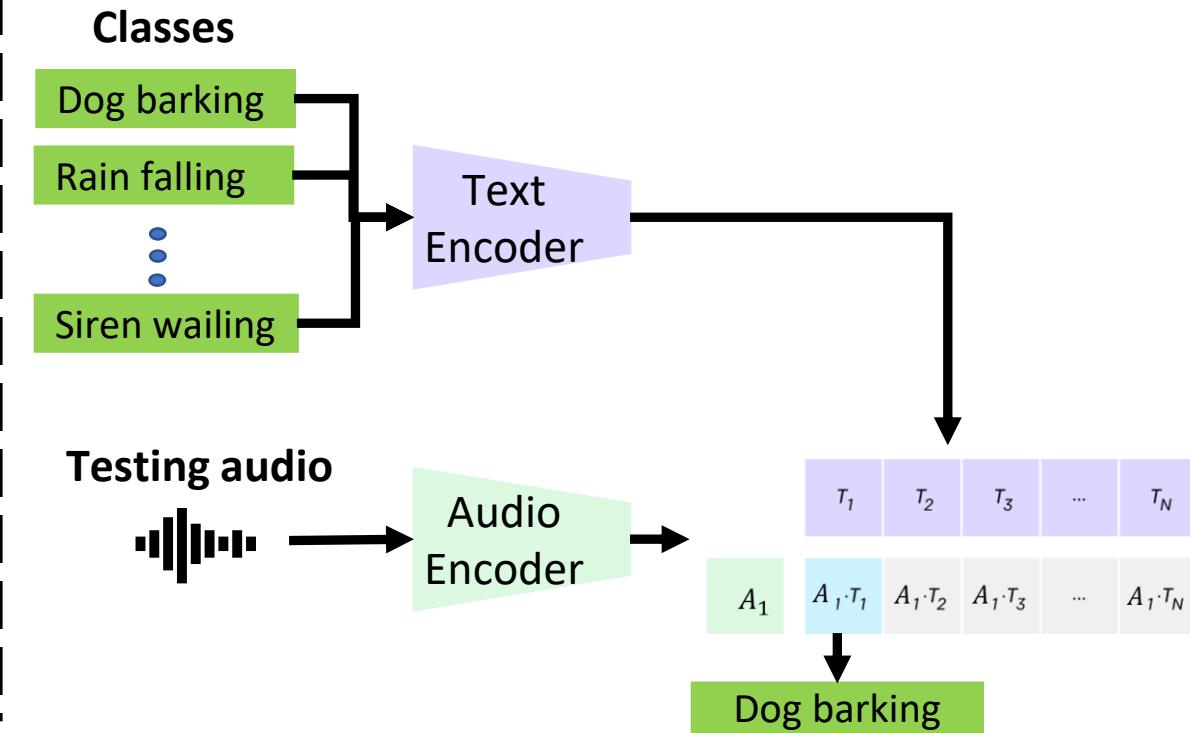
Move from classifier to embeddings and use similarity measure to determine prediction

# CLAP 🙌 Contrastive Language-Audio Pretraining

## 1. Contrastive pretraining

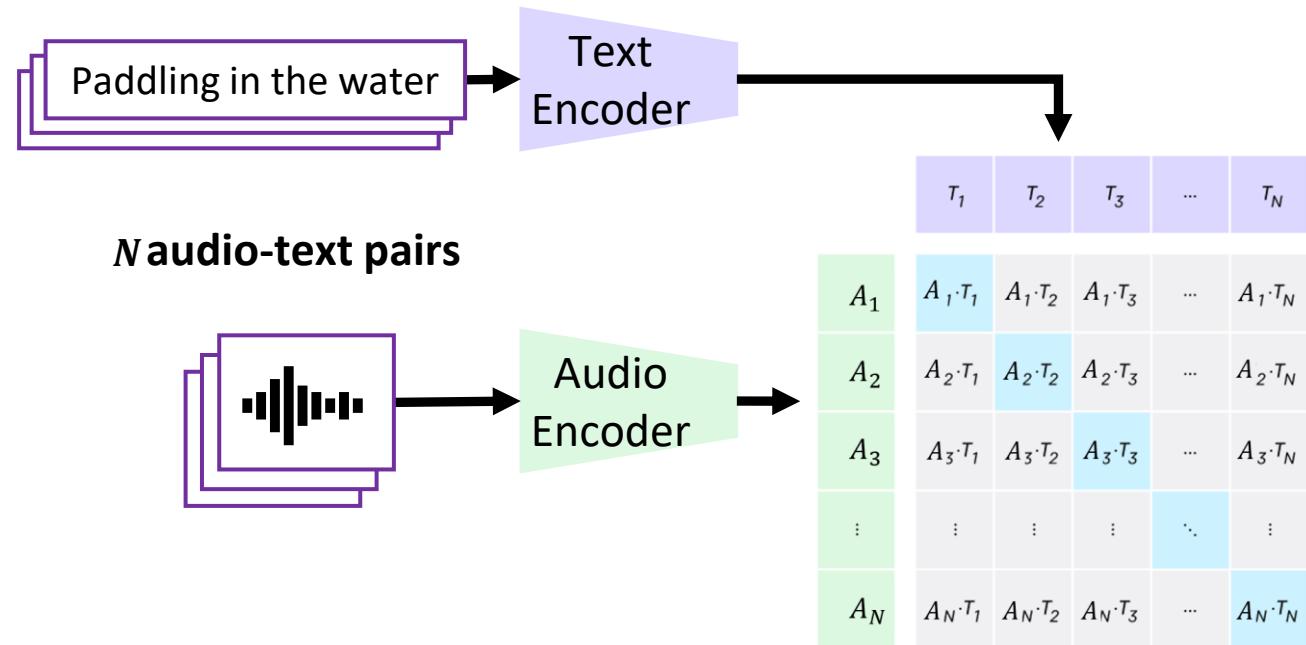


## 2. Zero-Shot classification



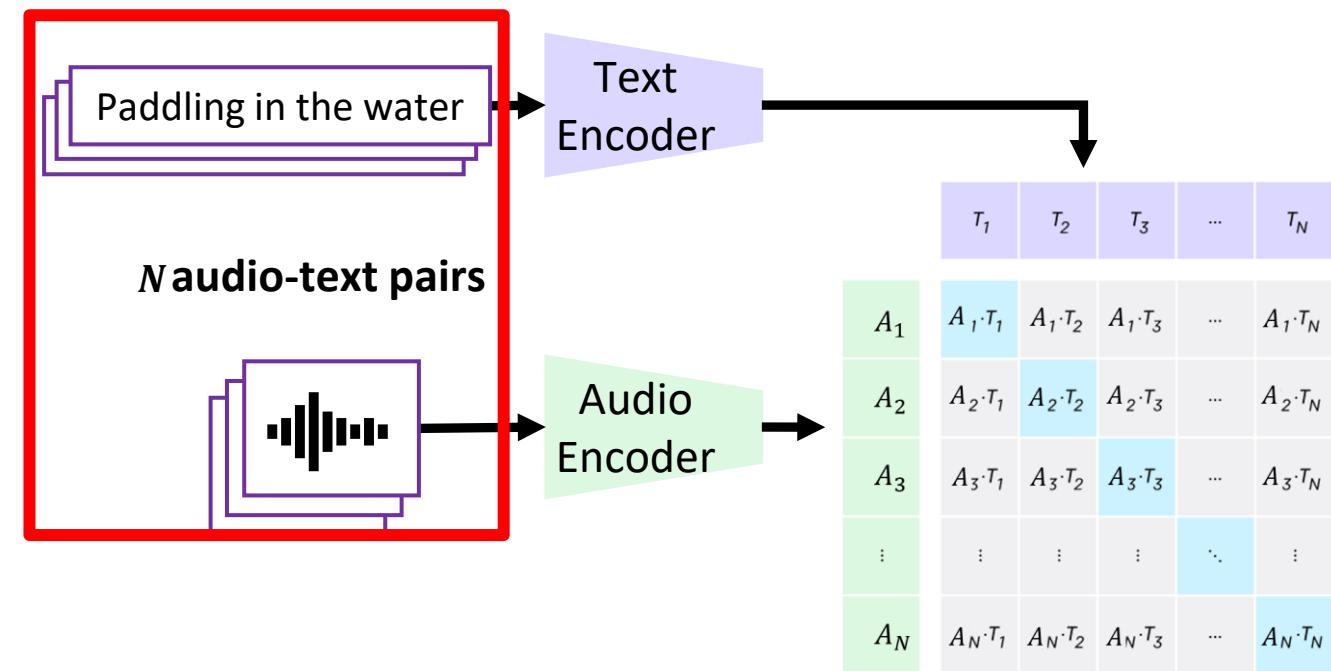
# Contrastive pretraining

## 1. Contrastive pretraining



# Contrastive pretraining

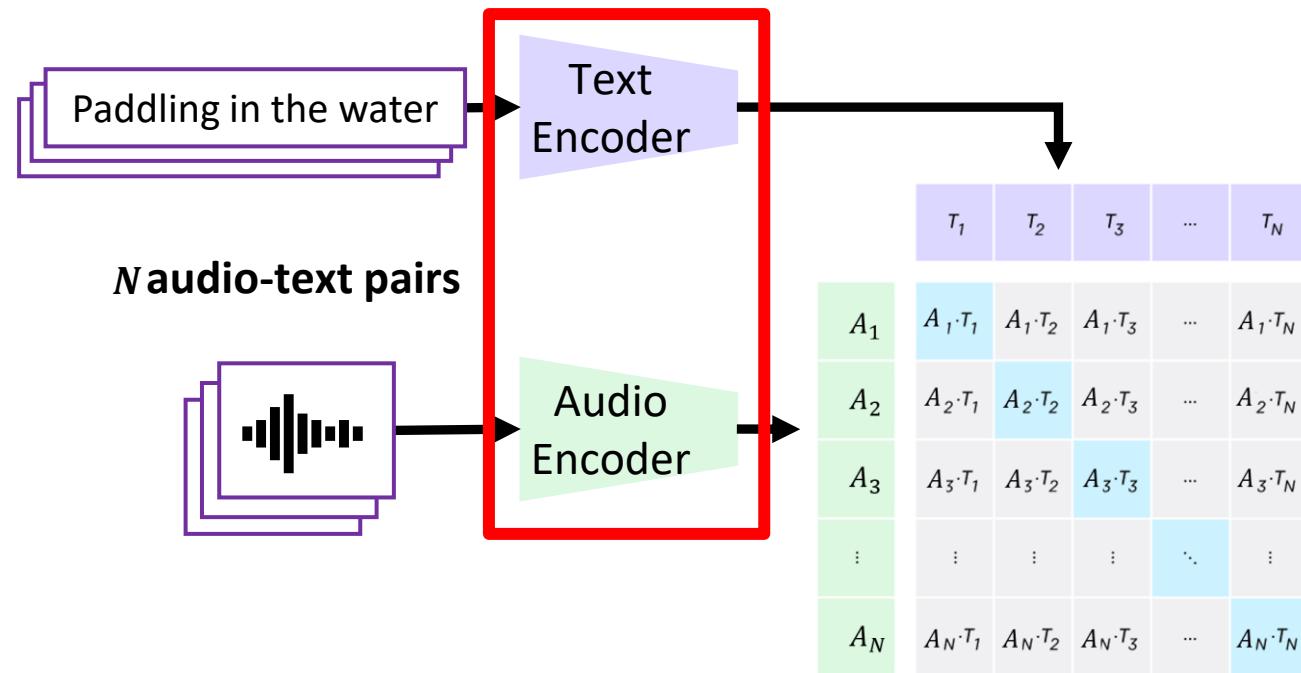
## 1. Contrastive pretraining



Training consists of batch of N audio-text pairs

# Contrastive pretraining

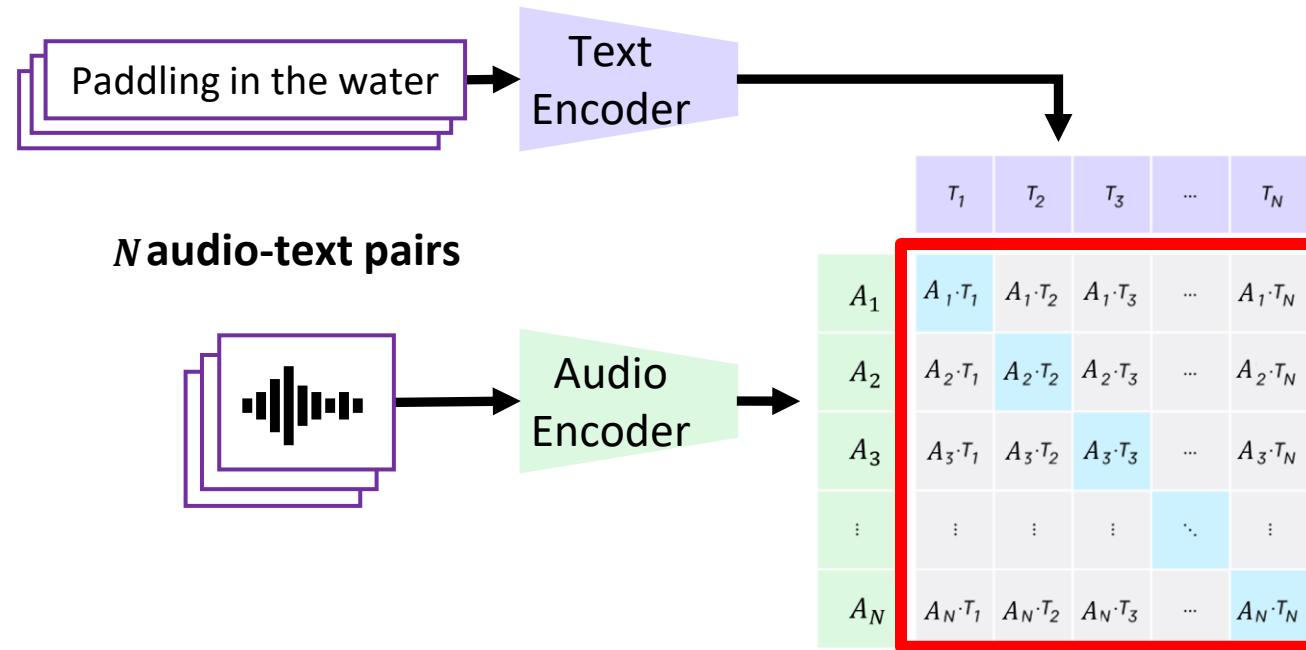
## 1. Contrastive pretraining



Independently encode audio-text pairs

# Contrastive pretraining

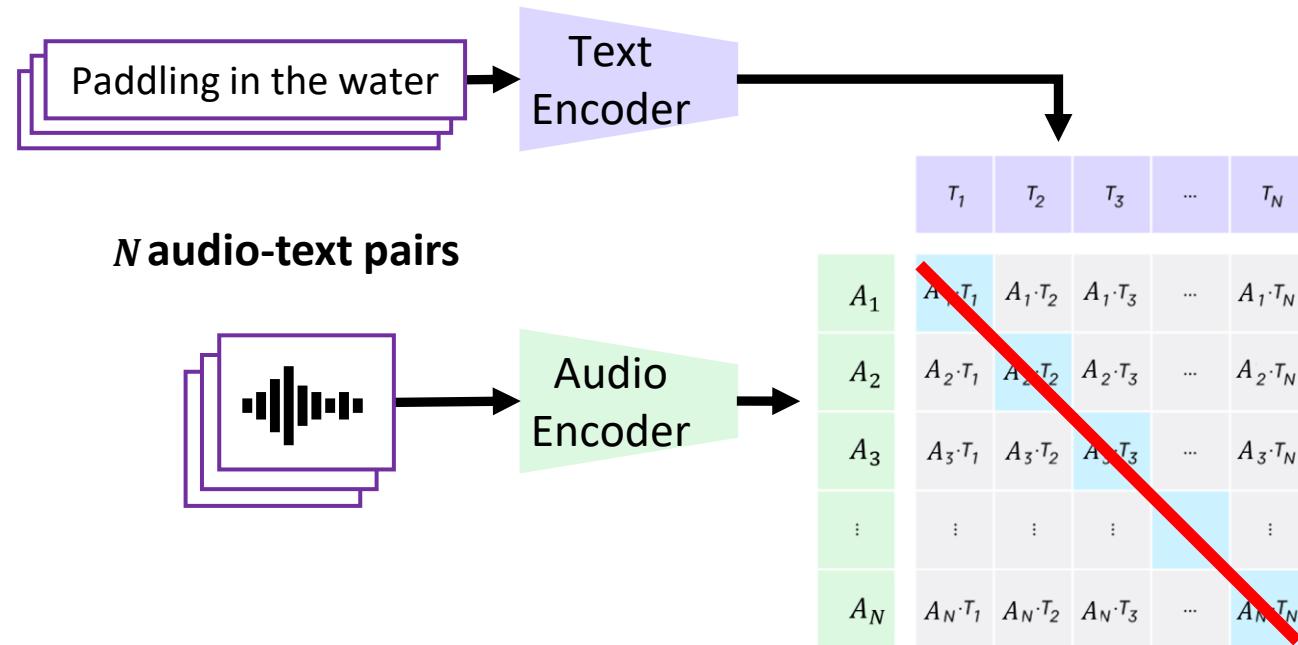
## 1. Contrastive pretraining



Compute dot product to form similarity matrix

# Contrastive pretraining

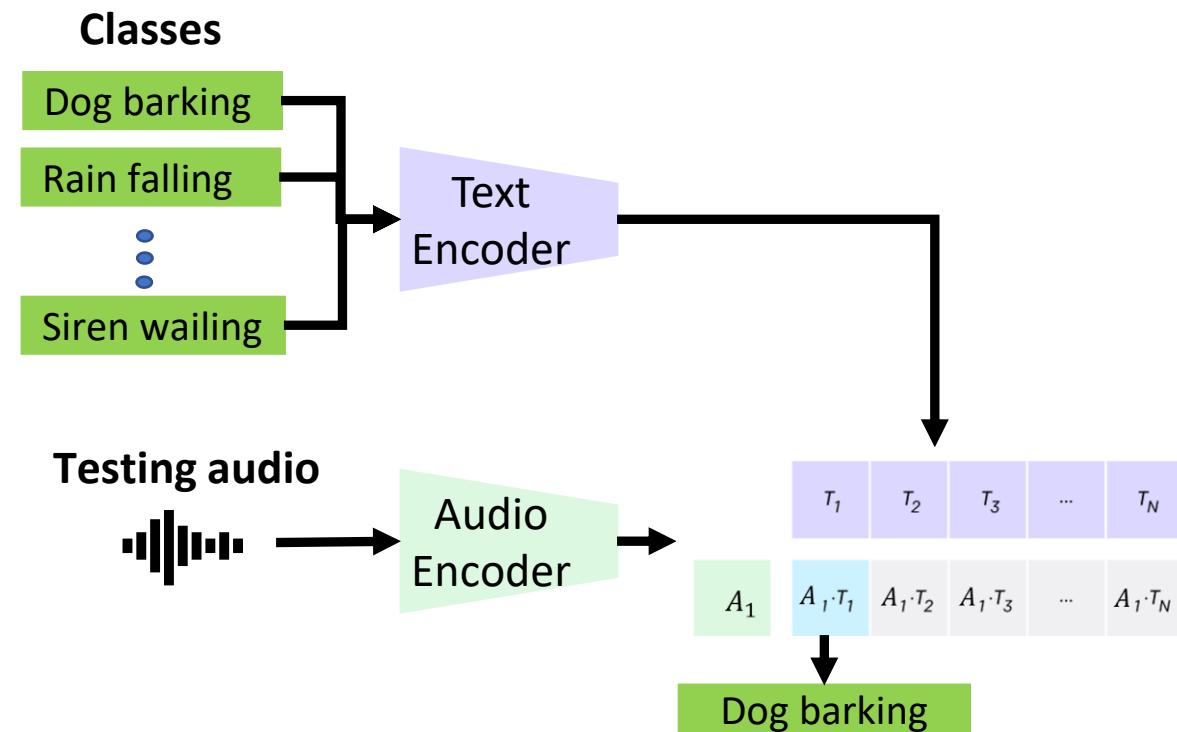
## 1. Contrastive pretraining



Minimize symmetric cross-entropy

# Zero-shot classification

## 2. Zero-Shot classification



# Training and evaluation dataset

- We use 128k audio-text pairs from 4 audio captioning datasets
- Some example captions

*The drum fill for when somebody tells a joke in a stand-up comic or a show*

*Soundscape taken at 1 am in Paris at the second-floor balcony of an apartment. Place Saint Augustin*

*Using a knife to cut zucchini on a wooden cutting board*

- We use 16 datasets from 8 different domains as downstream tasks for evaluation

# Training and evaluation dataset

- We use 128k audio-text pairs from 4 audio captioning datasets
- Some example captions

*The drum fill for when somebody tells a joke in a stand-up comic or a show*

***Soundscape taken at 1 am in Paris at the second-floor balcony of an apartment. Place Saint Augustin***

*Using a knife to cut zucchini on a wooden cutting board*

- We use 16 datasets from 8 different domains as downstream tasks for evaluation

# Zero-Shot classification results

Model	Sound Event Classification					Music			
	ESC50	FSD50K	US8K	DCASE17 Task 4	AudioSet	Music Speech	Music Genres	Mri. Stroke	Mri. Tonic
Random	0.02	< 0.005	0.1	0.05	< 0.0018	0.5	0.1	0.1	0.1667
Benchmark (ZS)	0.6940[10]	0.0302[9]	0.6531[10]	-	-	-	-	-	-
CLAP(ZS)	<b>0.826</b>	<b>0.3024</b>	<b>0.7324</b>	<b>0.3</b>	<b>0.058</b>	<b>1.0</b>	<b>0.252</b>	<b>0.3447</b>	<b>0.1965</b>

	Instrument Classification	Acoustic Scene Classification	Emotion Recognition		Keyword Spotting	Vocal Sound Classification	Speaker Counting
Model	Beijing Opera	TUT2017	CRE MA-D	RAV DESS	Speech Comm.	Vocal Sound	Libri Count
Random	0.25	0.06	0.1667	0.125	0.083	0.1667	0.090
CLAP (ZS)	<b>0.4746</b>	<b>0.2963</b>	<b>0.1784</b>	<b>0.1599</b>	<b>0.1063</b>	<b>0.4945</b>	<b>0.1788</b>

**Table 1.** CLAP (ZS) Zero-Shot outperforms the literature.

Higher is better for all numbers, DCASE17 employs F1, FSD50K and AudioSet employs mAP, everything else uses accuracy.

# Zero-Shot classification results

Model	Sound Event Classification					Music			
	ESC50	FSD50K	US8K	DCASE17 Task 4	AudioSet	Music Speech	Music Genres	Mri. Stroke	Mri. Tonic
Random	0.02	< 0.005	0.1	0.05	< 0.0018	0.5	0.1	0.1	0.1667
Benchmark (ZS)	0.6940[10]	0.0302[9]	0.6531[10]	-	-	-	-	-	-
CLAP(ZS)	<b>0.826</b>	<b>0.3024</b>	<b>0.7324</b>	<b>0.3</b>	<b>0.058</b>	<b>1.0</b>	<b>0.252</b>	<b>0.3447</b>	<b>0.1965</b>

Model	Instrument Classification	Acoustic Scene Classification	Emotion Recognition		Keyword Spotting	Vocal Sound Classification	Speaker Counting
	Beijing Opera	TUT2017	CRE MA-D	RAV DESS	Speech Comm.	Vocal Sound	Libri Count
Random	0.25	0.06	0.1667	0.125	0.083	0.1667	0.090
CLAP (ZS)	<b>0.4746</b>	<b>0.2963</b>	<b>0.1784</b>	<b>0.1599</b>	<b>0.1063</b>	<b>0.4945</b>	<b>0.1788</b>

**Table 1.** CLAP (ZS) Zero-Shot outperforms the literature.

Higher is better for all numbers, DCASE17 employs F1, FSD50K and AudioSet employs mAP, everything else uses accuracy.

# Zero-Shot classification results

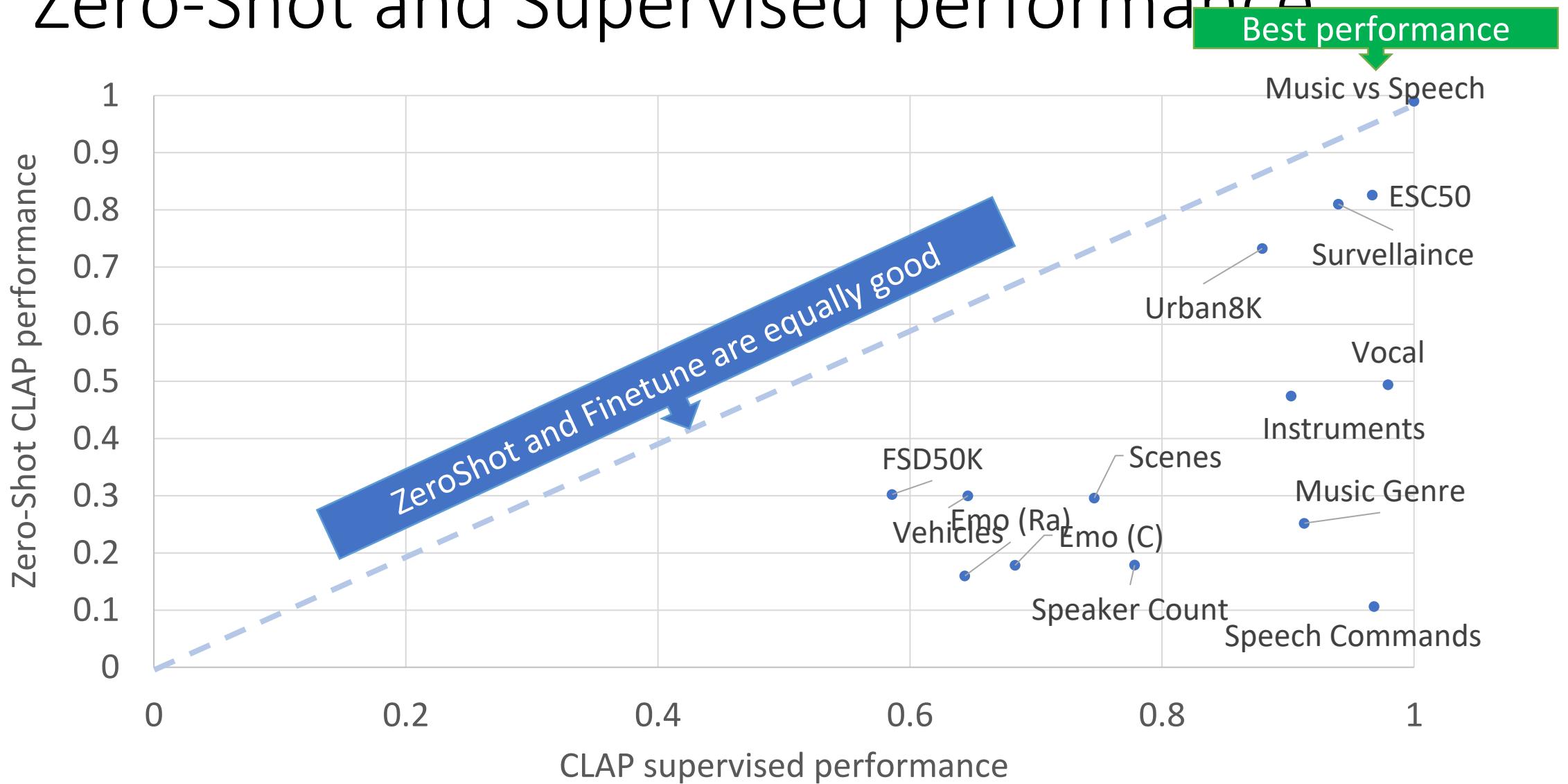
Model	Sound Event Classification					Music			
	ESC50	FSD50K	US8K	DCASE17 Task 4	AudioSet	Music Speech	Music Genres	Mri. Stroke	Mri. Tonic
Random Benchmark (ZS)	0.02 0.6940[10]	< 0.005 0.0302[9]	0.1 0.6531[10]	0.05 -	< 0.0018 -	0.5 -	0.1 -	0.1 -	0.1667 -
CLAP(ZS)	<b>0.826</b>	<b>0.3024</b>	<b>0.7324</b>	<b>0.3</b>	<b>0.058</b>	<b>1.0</b>	<b>0.252</b>	<b>0.3447</b>	<b>0.1965</b>

	Instrument Classification	Acoustic Scene Classification	Emotion Recognition		Keyword Spotting	Vocal Sound Classification	Speaker Counting
Model	Beijing Opera	TUT2017	CRE MA-D	RAV DESS	Speech Comm.	Vocal Sound	Libri Count
Random	0.25	0.06	0.1667	0.125	0.083	0.1667	0.090
CLAP (ZS)	<b>0.4746</b>	<b>0.2963</b>	<b>0.1784</b>	<b>0.1599</b>	<b>0.1063</b>	<b>0.4945</b>	<b>0.1788</b>

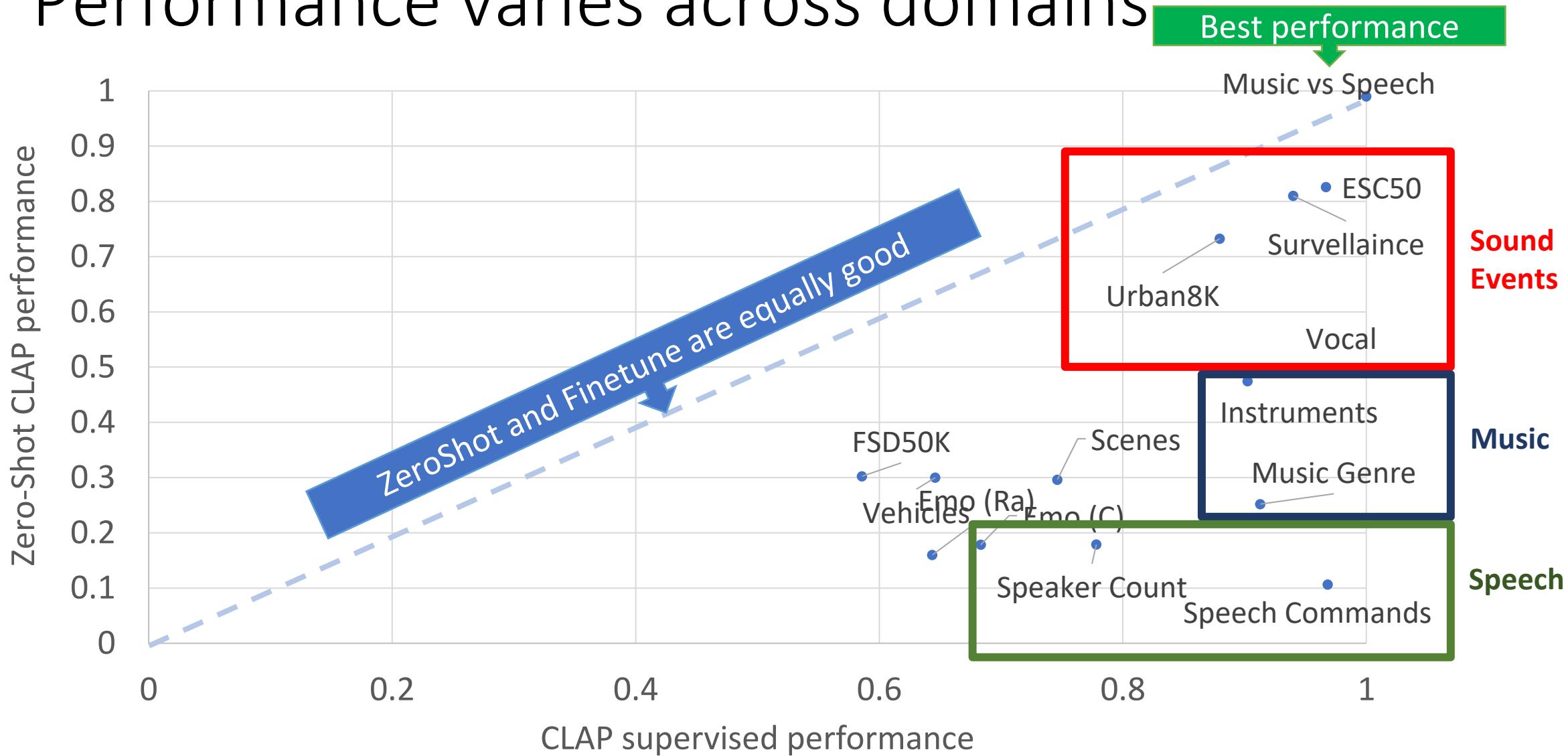
**Table 1.** CLAP (ZS) Zero-Shot outperforms the literature.

Higher is better for all numbers, DCASE17 employs F1, FSD50K and AudioSet employs mAP, everything else uses accuracy.

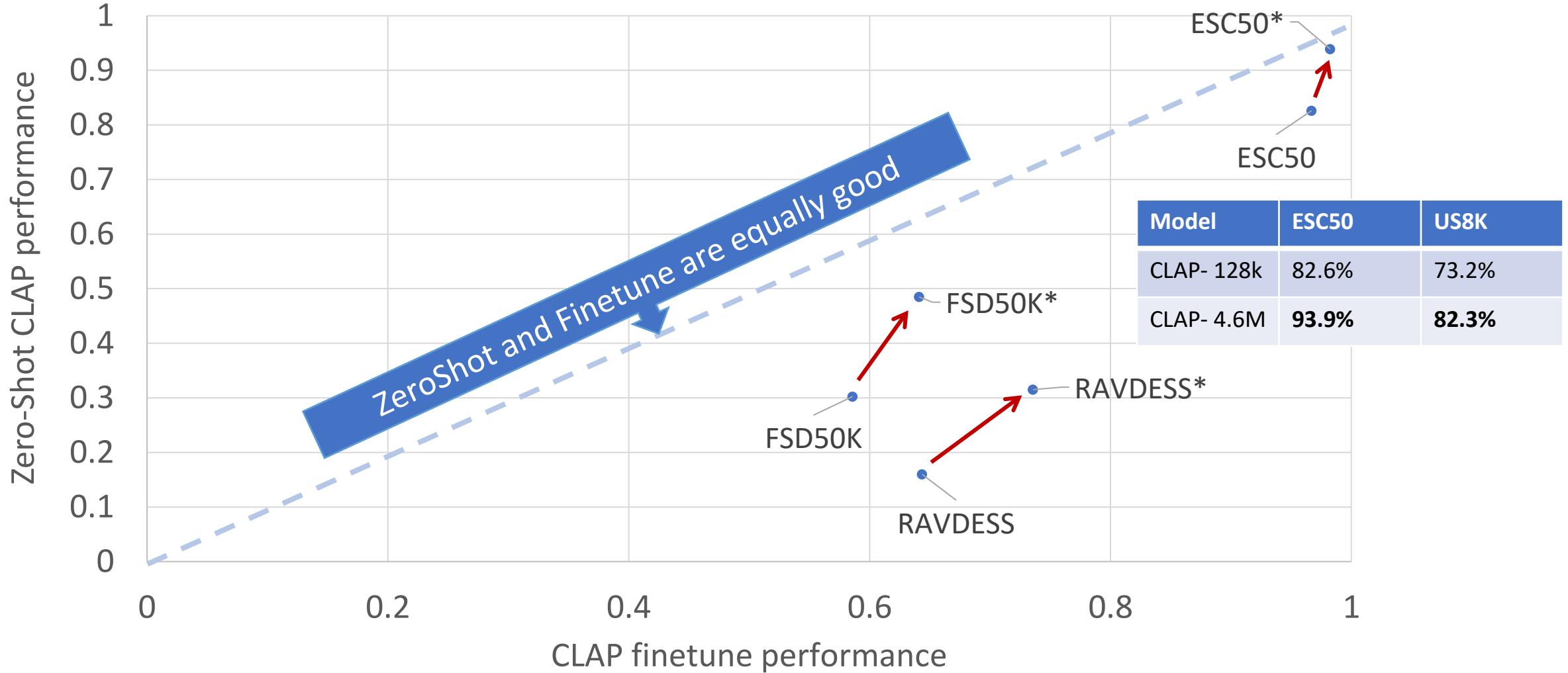
# Zero-Shot and Supervised performance



# Performance varies across domains



# Scaling reduces gap in ZS and supervised



# Contrastive pretraining

Can be used for close-ended tasks, such as classification and retrieval

However, inherently lack the capacity to perform open-ended tasks, such as Audio Captioning or Audio Question & Answering

# Contrastive pretraining

Can be used for close-ended tasks, such as classification and retrieval

However, inherently lack the capacity to perform open-ended tasks, such as Audio Captioning or Audio Question & Answering

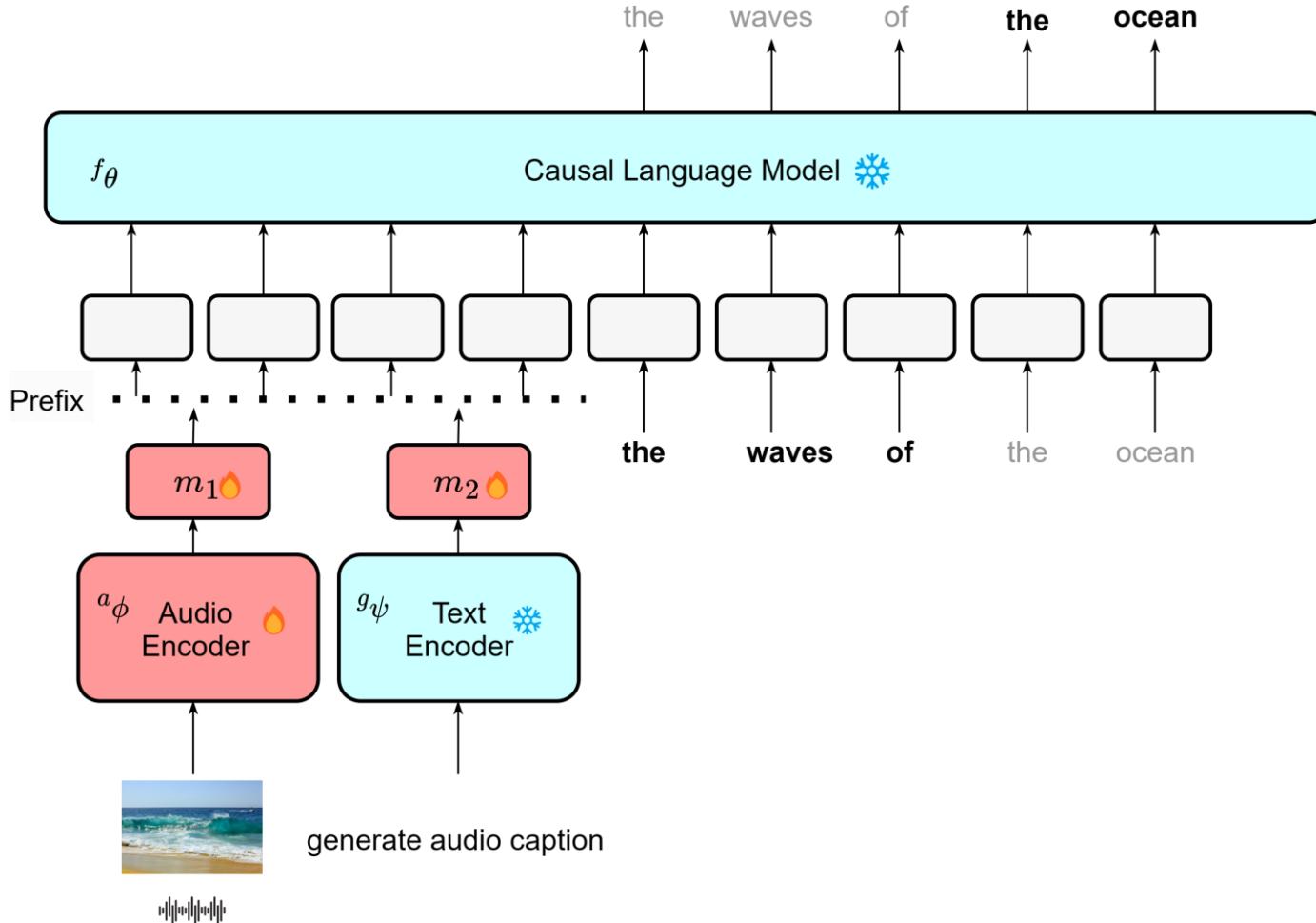
**A unified model for both close-ended and open-ended tasks?**



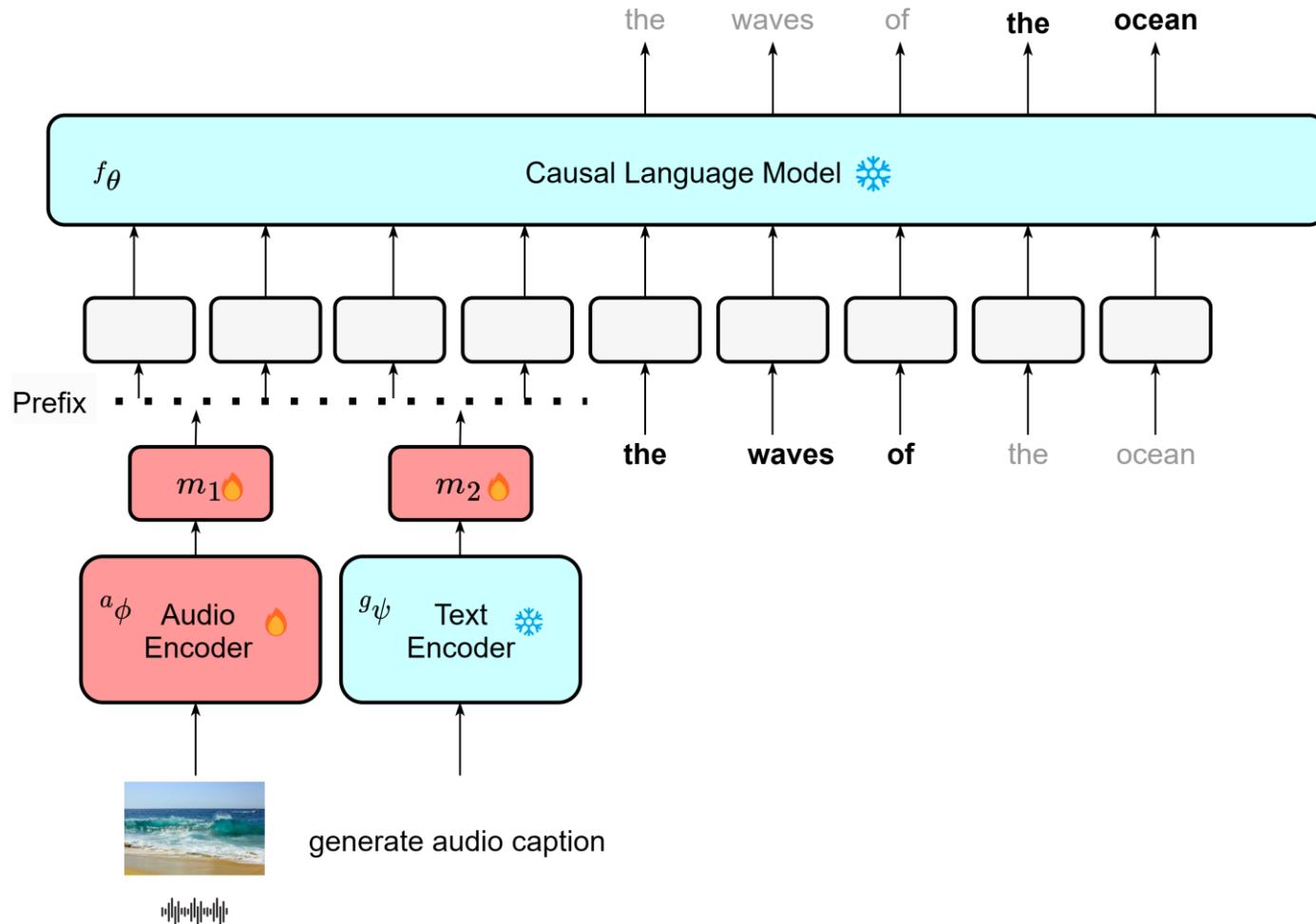
# Pengi: An Audio Language Model

Audio input	Text input	Text output
	 this is a sound of	train, railway and locomotive
	 this emotion is	happy
	 generate audio caption	the waves of the ocean crash onto the shore then recede
	 question: what type of animal is making the light sound in the background?	it is a bird

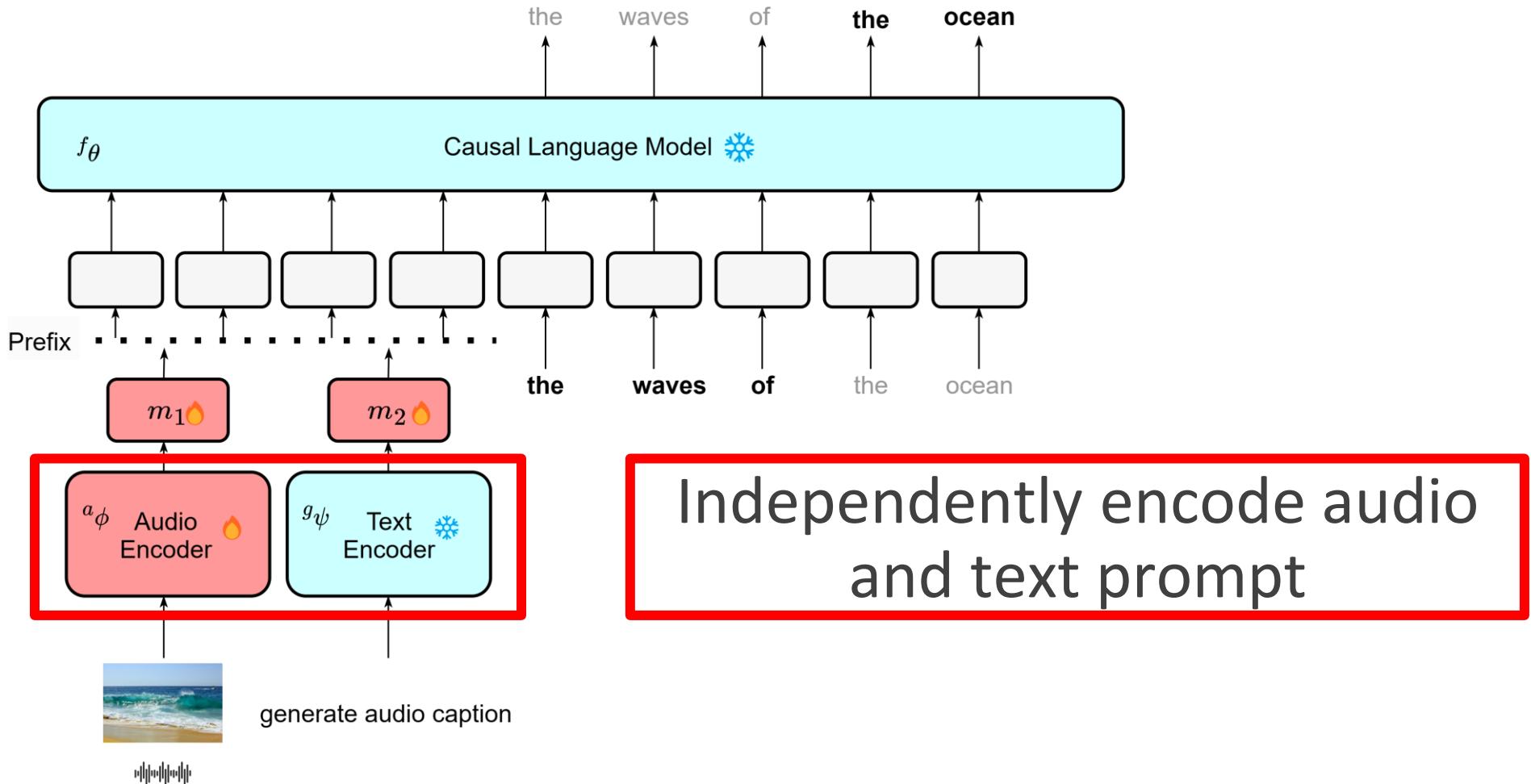
# Model architecture



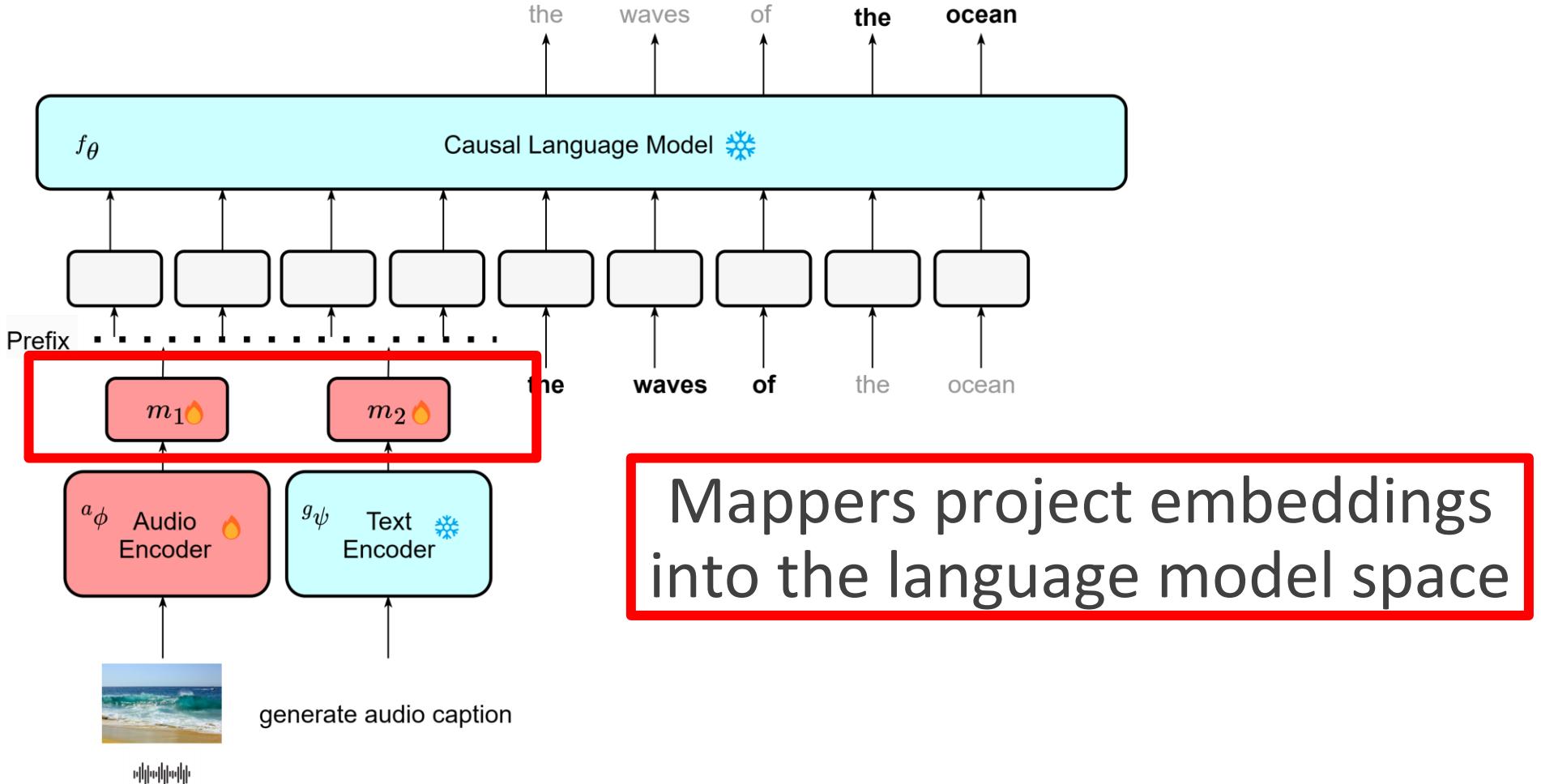
# Training



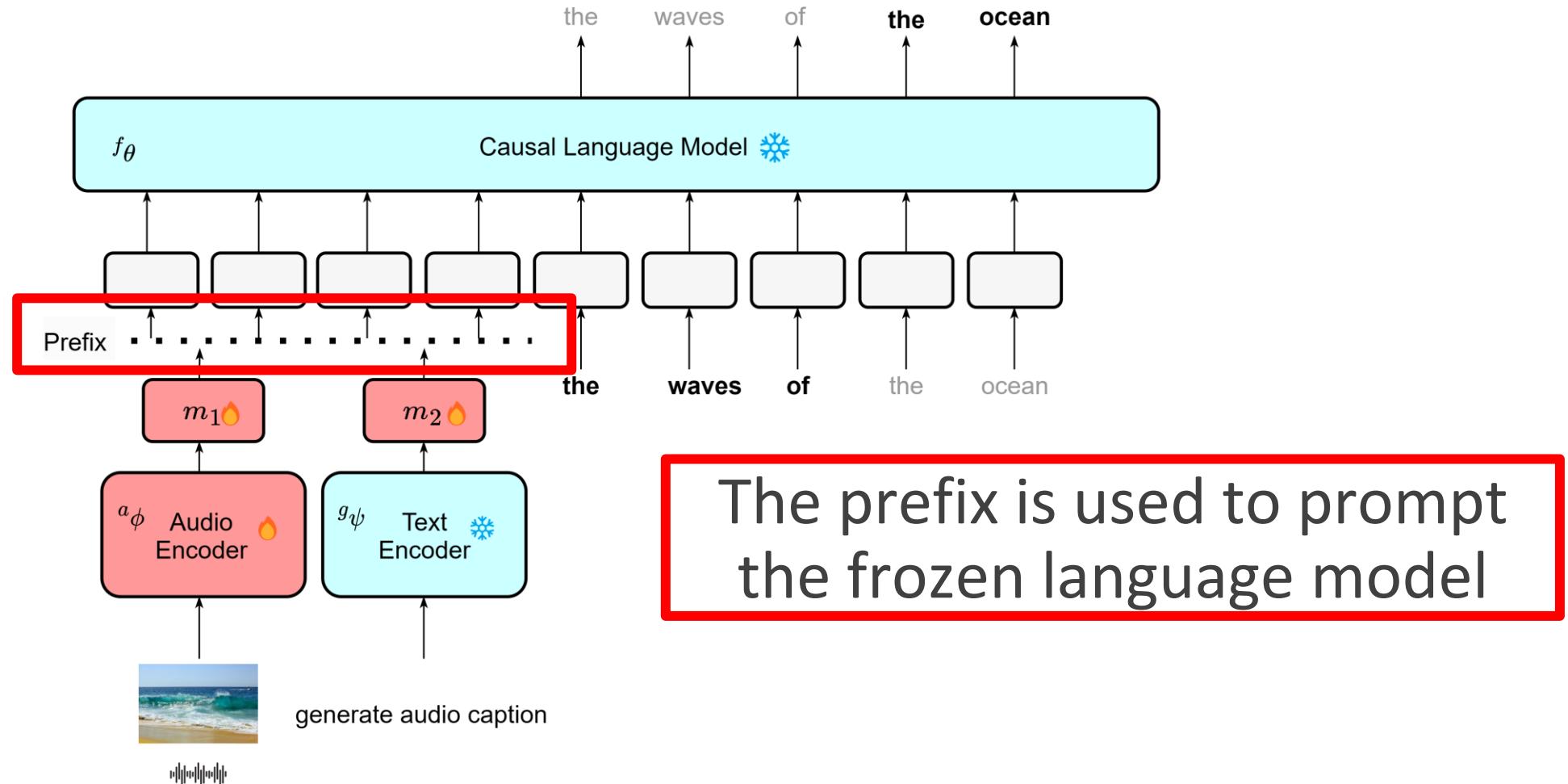
# Training



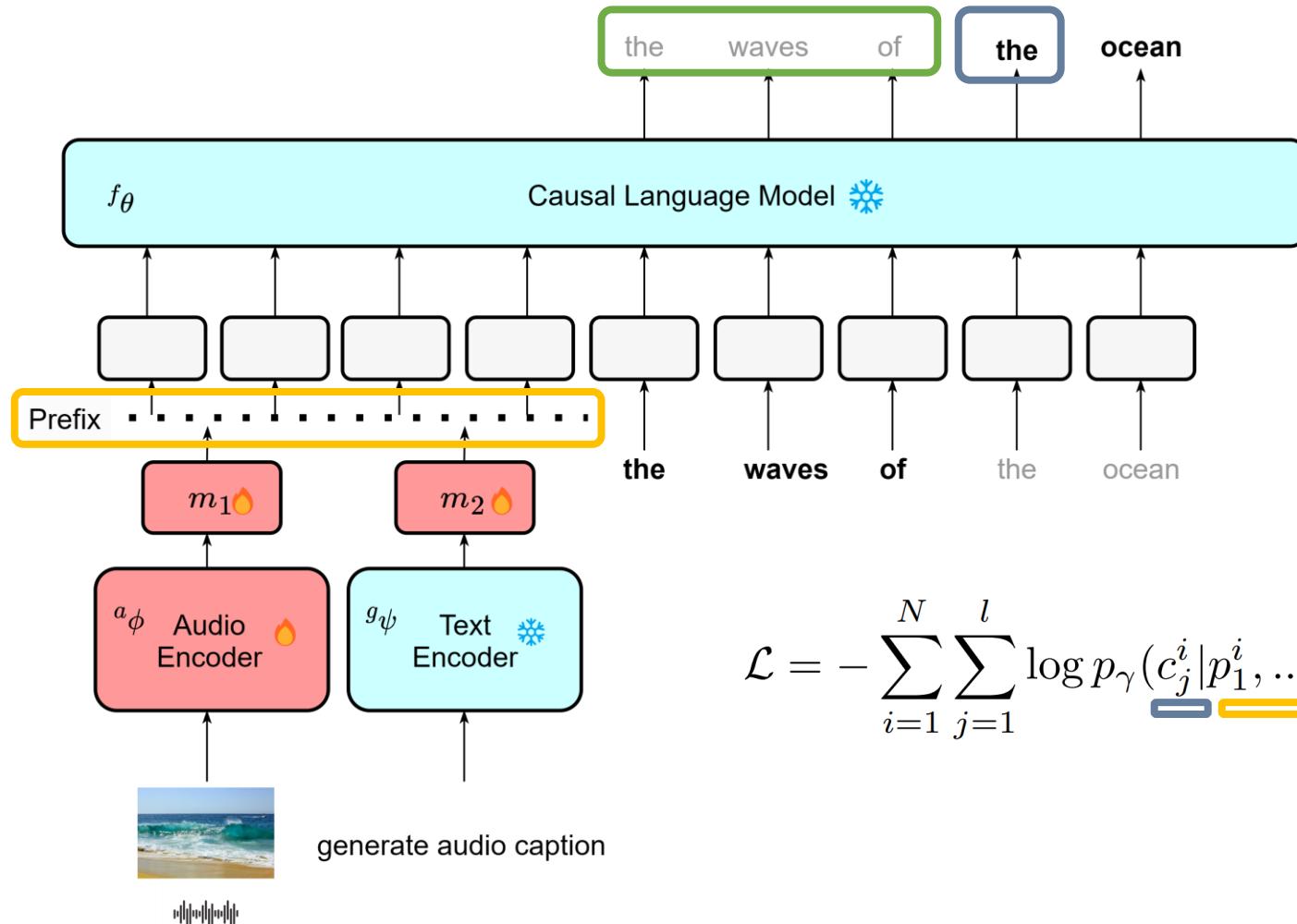
# Training



# Training

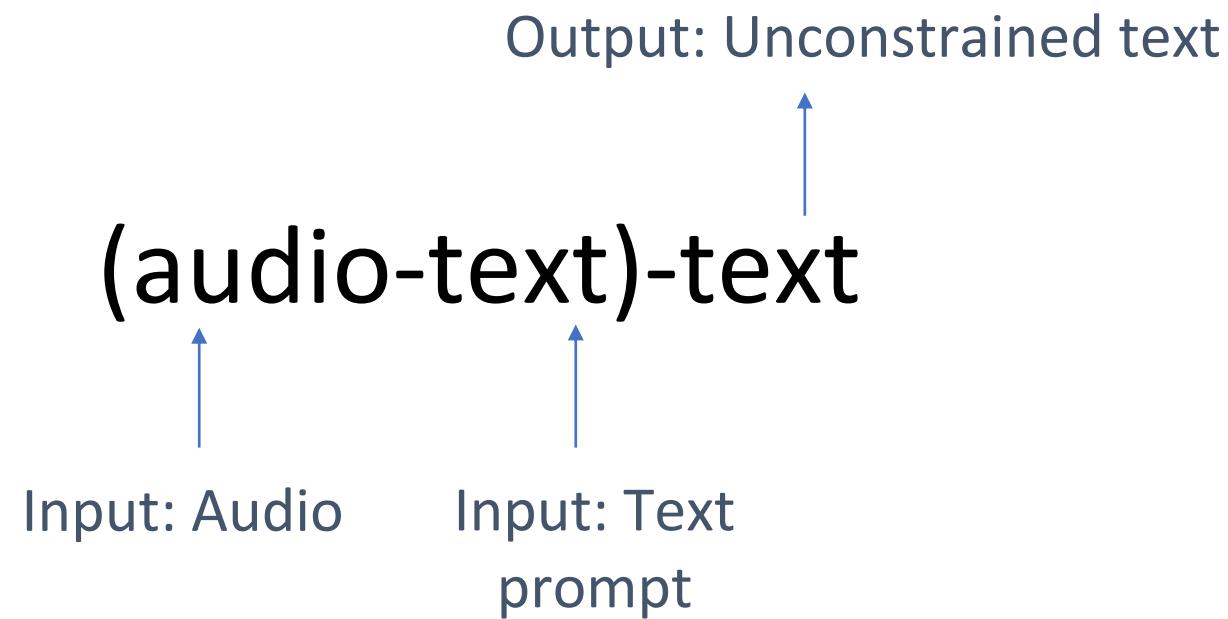


# Training



$$\mathcal{L} = - \sum_{i=1}^N \sum_{j=1}^l \log p_\gamma(c_j^i | p_1^i, \dots, p_{2k}^i, c_1^i, \dots, c_{j-1}^i)$$

# Audio-task templates for training



**3.4M triplets**

# Benchmarking Pungi on downstream tasks

Model	Audio Captioning ↑		Audio Q&A ↑		Sound Event Classification ↑			
	AudioCaps	Clotho	ClothoAQA	ESC50	FSD50K	US8K	DCASE17 Task 4	
CLAP*	X	X	X	0.8916	0.3398	<b>0.7661</b>	<b>0.3387</b>	
Pungi	<b>0.4667</b>	<b>0.2709</b>	<b>0.6453</b>	<b>0.9195</b>	<b>0.4676</b>	0.7185	0.3380	
Model	Acoustic Scene Classification↑		Music ↑		Instrument Classification ↑		Music Note Analysis↑	
	TUT2017	Music Speech	Music Genres	Beijing Opera	Instrument family	NS. Pitch	NS. Velocity	NS. Qualities
CLAP*	0.3037	<b>1.0</b>	<b>0.479</b>	0.4025	0.415	0.1337	0.2185	0.2545
Pungi	<b>0.3525</b>	0.9688	0.3525	<b>0.6229</b>	<b>0.5007</b>	<b>0.8676</b>	<b>0.3728</b>	<b>0.386</b>
Model	Emotion Recognition↑			Vocal Sound Classification↑	Action Recog.↑	Survei llance.↑		
	CRE	RAV	DESS	Vocal Sound	ESC50 Actions	SESA		
CLAP*	0.1512	0.1692		0.5522	0.508	<b>0.7094</b>		
Pungi	<b>0.1846</b>	<b>0.2032</b>		<b>0.6035</b>	<b>0.5277</b>	0.5402		

# Benchmarking Pungi on downstream tasks

Model	Audio Captioning ↑		Audio Q&A ↑		Sound Event Classification ↑			
	AudioCaps	Clotho	ClothoAQA	ESC50	FSD50K	US8K	DCASE17 Task 4	
CLAP*	X	X	X	0.8916	0.3398	<b>0.7661</b>	<b>0.3387</b>	
Pungi	<b>0.4667</b>	<b>0.2709</b>	<b>0.6453</b>	<b>0.9195</b>	<b>0.4676</b>	0.7185	0.3380	
Acoustic Scene Classification↑		Music ↑		Instrument Classification ↑		Music Note Analysis↑		
Model	TUT2017	Music Speech	Music Genres	Beijing Opera	Instrument family	NS. Pitch	NS. Velocity	NS. Qualities
CLAP*	0.3037	<b>1.0</b>	<b>0.479</b>	0.4025	0.415	0.1337	0.2185	0.2545
Pungi	<b>0.3525</b>	0.9688	0.3525	<b>0.6229</b>	<b>0.5007</b>	<b>0.8676</b>	<b>0.3728</b>	<b>0.386</b>
		Emotion Recognition↑		Vocal Sound Classification↑	Action Recog.↑	Survei llance.↑		
Model	CRE MA-D	RAV DESS	Vocal Sound	ESC50 Actions	SESA			
CLAP*	0.1512	0.1692	0.5522	0.508	<b>0.7094</b>			
Pungi	<b>0.1846</b>	<b>0.2032</b>	<b>0.6035</b>	<b>0.5277</b>	0.5402			

# Benchmarking Pungi on downstream tasks

	Audio Captioning ↑		Audio Q&A ↑		Sound Event Classification ↑			
Model	AudioCaps	Clotho	ClothoAQA		ESC50	FSD50K	US8K	DCASE17 Task 4
CLAP*	X	X		X	0.8916	0.3398	<b>0.7661</b>	<b>0.3387</b>
Pungi	<b>0.4667</b>	<b>0.2709</b>		<b>0.6453</b>	<b>0.9195</b>	<b>0.4676</b>	0.7185	0.3380

	Acoustic Scene Classification↑	Music ↑		Instrument Classification ↑		Music Note Analysis↑		
Model	TUT2017	Music Speech	Music Genres	Beijing Opera	Instrument family	NS. Pitch	NS. Velocity	NS. Qualities
CLAP*	0.3037	<b>1.0</b>	<b>0.479</b>	0.4025	0.415	0.1337	0.2185	0.2545
Pungi	<b>0.3525</b>	0.9688	0.3525	<b>0.6229</b>	<b>0.5007</b>	<b>0.8676</b>	<b>0.3728</b>	<b>0.386</b>

	Emotion Recognition↑		Vocal Sound Classification↑	Action Recog.↑	Survei llance.↑
Model	CRE MA-D	RAV DESS	Vocal Sound	ESC50 Actions	SESA
CLAP*	0.1512	0.1692	0.5522	0.508	<b>0.7094</b>
Pungi	<b>0.1846</b>	<b>0.2032</b>	<b>0.6035</b>	<b>0.5277</b>	0.5402

# Generative pretraining

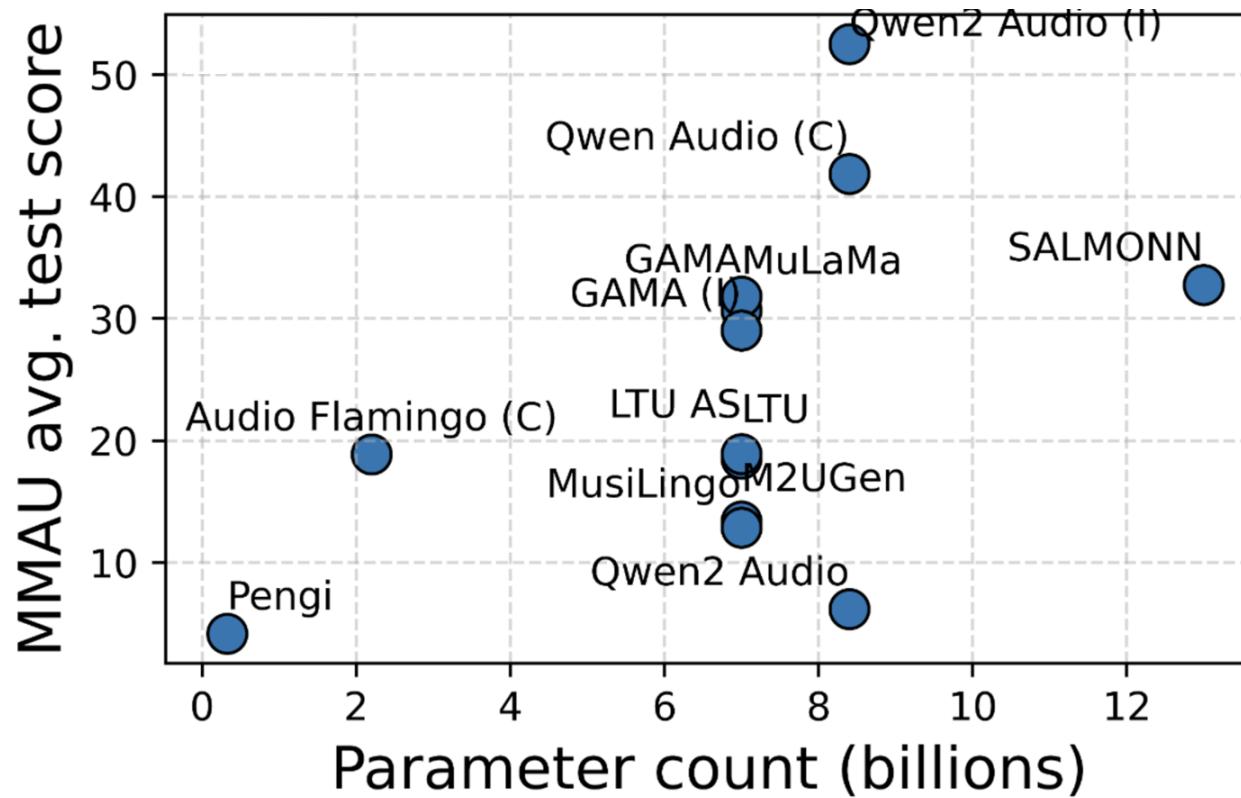
Pengi allows the user to perform both open-ended tasks, such as audio captioning and question-answering, and close-ended tasks like audio classification and retrieval

# Generative pretraining

Pengi allows the user to perform both open-ended tasks, such as audio captioning and question-answering, and close-ended tasks like audio classification and retrieval

**However, reasoning performance of Pengi is still near random**

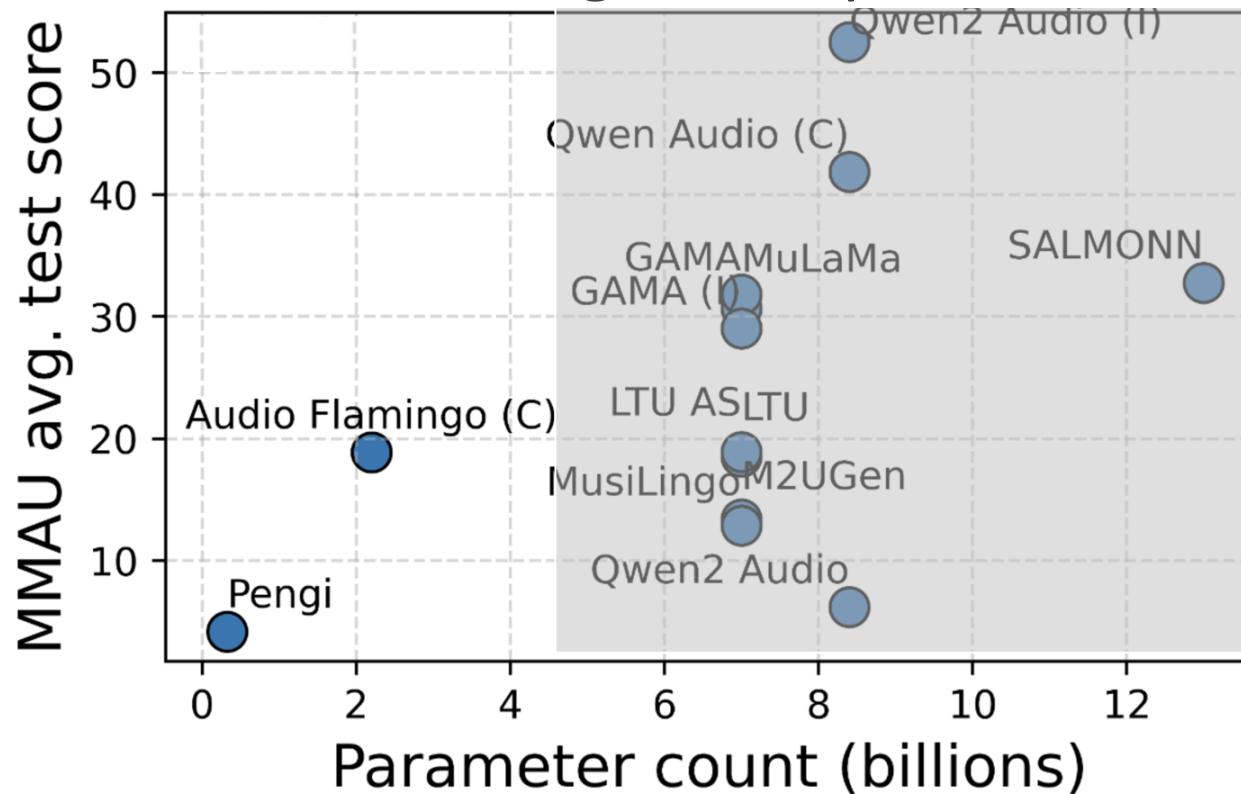
# Analyzing the reasoning performance of AFM



MMAU: A Massive Multi-Task Audio Understanding and Reasoning Benchmark, <https://arxiv.org/abs/2410.19168>, ICLR 2025

# Analyzing the reasoning performance of AFM

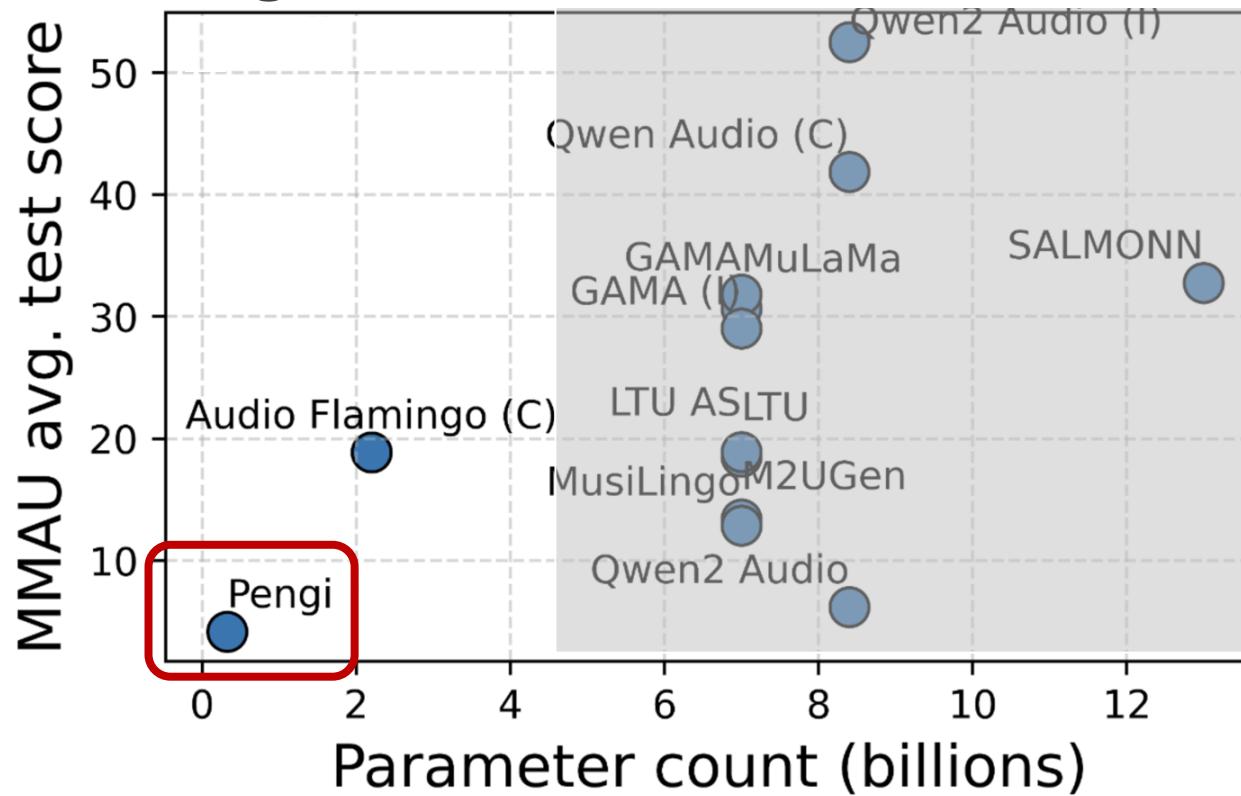
AFMs start to exhibit reasoning at ~8B parameters and 3M+ pairs



MMAU: A Massive Multi-Task Audio Understanding and Reasoning Benchmark, <https://arxiv.org/abs/2410.19168>, ICLR 2025

# Analyzing the reasoning performance of AFM

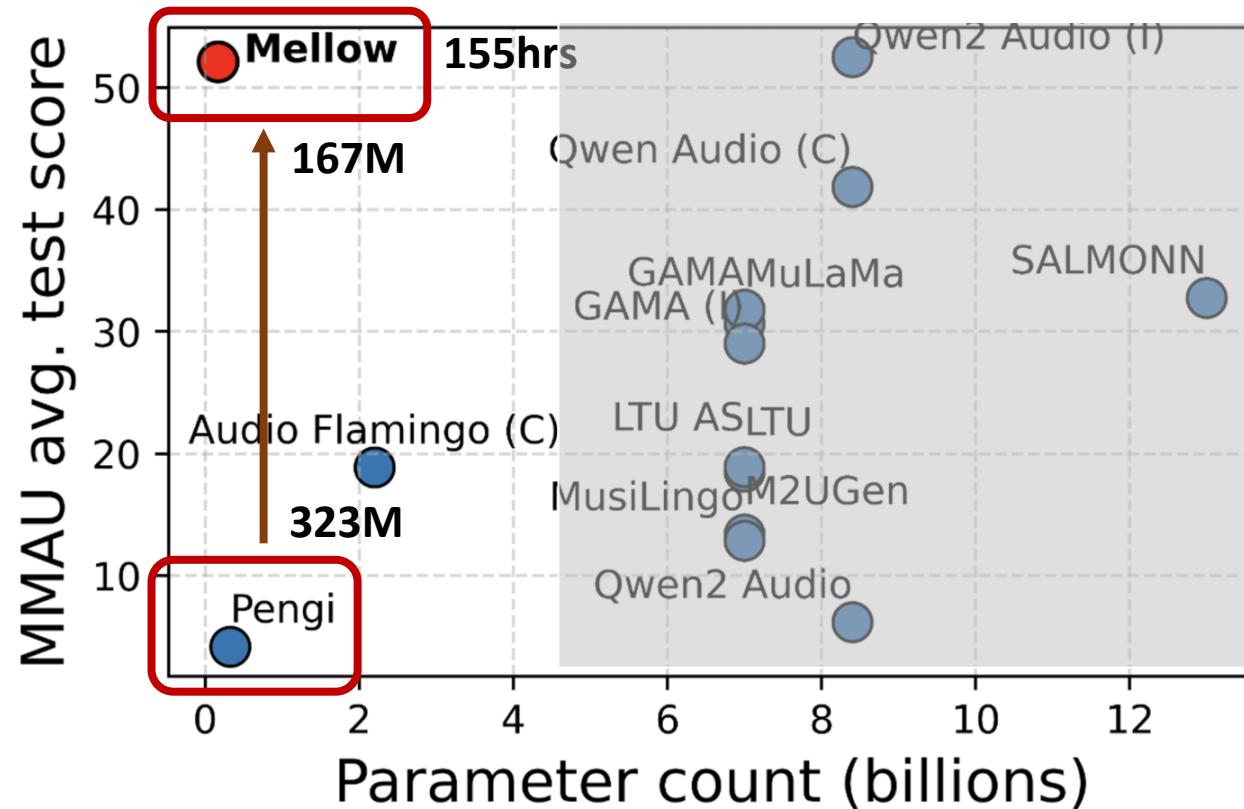
Can we enable Pengi-like models to do audio reasoning?



MMAU: A Massive Multi-Task Audio Understanding and Reasoning Benchmark, <https://arxiv.org/abs/2410.19168>, ICLR 2025

# Mellow: small audio-language model for reasoning

SoTA performance using 167M parameters and 155 hrs of aligned training



Mellow: a small audio-language model for reasoning,  
<https://arxiv.org/abs/2503.08540>, under review

# Mellow: small audio-language model for reasoning

## Q: What can be inferred about the materials involved in the sound?

M: From the sounds, it can be inferred that the toilet bowl and pipe are likely made of ceramic or porcelain, as these materials produce a distinctive sound when water flows through them. The sound of the water draining also suggests that the pipe is likely made of a smooth material, such as PVC or ABS, which produces a low-frequency rumble when water flows through it.

## Q: How do the two sound events interact with each other?

M: The baby's crying sound and the TV sound interact with each other in a way that the crying sound dominates the auditory scene. The TV sound is present in the background, but it is masked by the loudness of the crying sound. This suggests that the caregiver or family member is trying to create a distraction or background noise to soothe the baby.

## Q: Explain the difference between the two audios in detail

M: As I listen to the two audios, I notice distinct differences in their acoustic properties. The first audio, featuring a kid screaming at a park, has a sharp, high-pitched frequency with a rising loudness level, indicating a sudden emotional outburst. The background noise of people chatting and laughing creates a gentle hum, adding a sense of an environment context. In contrast, the second audio is a truck driving down the street, has a deep, rumbling frequency with a steady, constant loudness level, characteristic of a powerful engine. The background noise of traffic, including car horns and revving engines, creates a cacophonous soundscape, evoking a sense of urban chaos. While both audios have high levels of loudness, the kid's scream is more attention-grabbing, whereas the truck's engine is sustained.

## Q: What is the characteristic of the machine's sound? A) Whiny and squealing B) High-pitched and smooth C) Low-pitched and rough D) Silent

M: A) Whiny and squealing

## Q: What can be inferred about the environment?

M: The environment is likely a natural setting, such as a pond or lake, with a body of water, vegetation, and possibly some buildings or structures in the background, based on the sounds of water splashing and rustling.

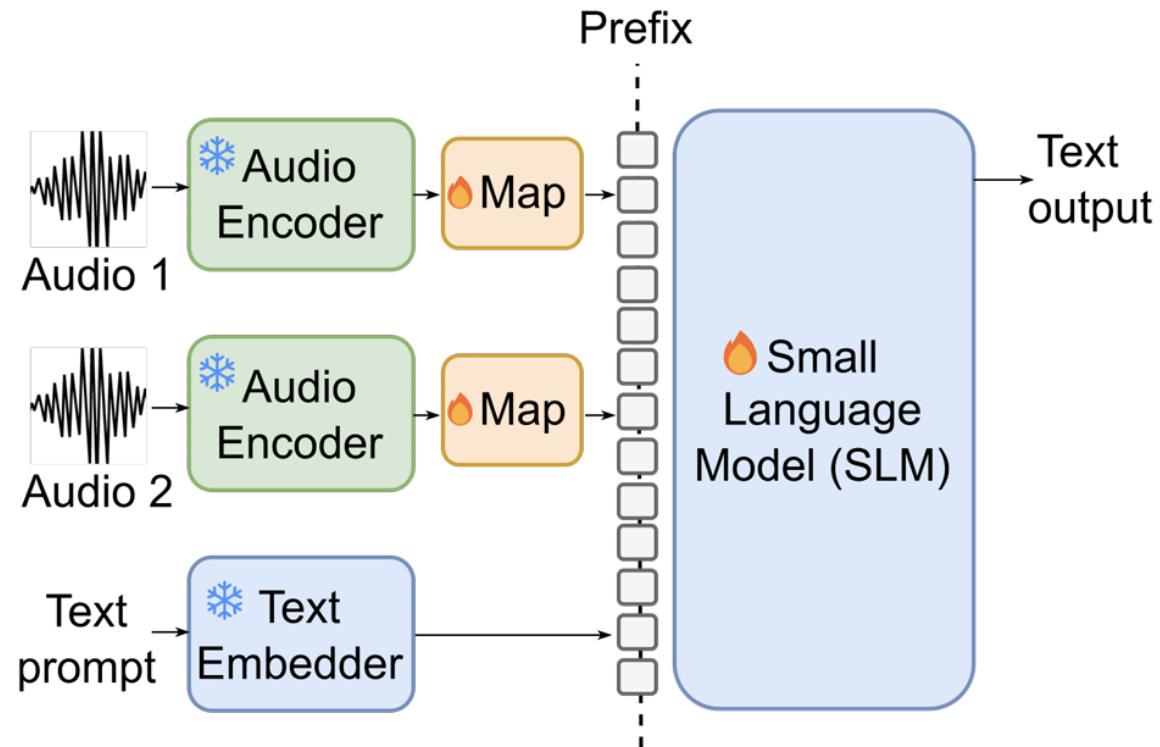
## Q: How do the duck calls relate to the other sound events in the clip?

M: The duck calls are likely being blown by a person, and are used to interact with the ducks, possibly to lure them or to communicate with them, and are heard in conjunction with the human conversation, suggesting a scene of duck hunting or observation.

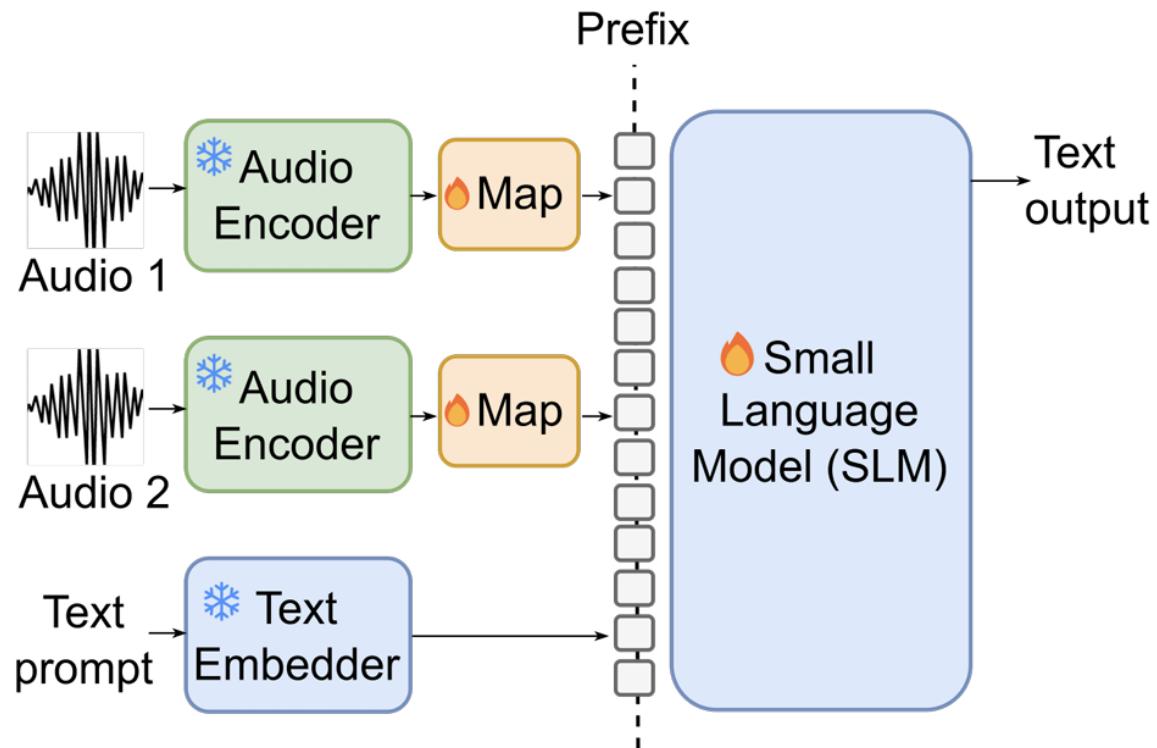
## Q: Based on the audio, what can be said about the hypothesis - "A farmer is giving a tour of his ranch while chickens roam nearby"? a) It is definitely true b) It is definitely false c) It is plausible d) I cannot determine

M: b) It is definitely false

# Mellow architecture



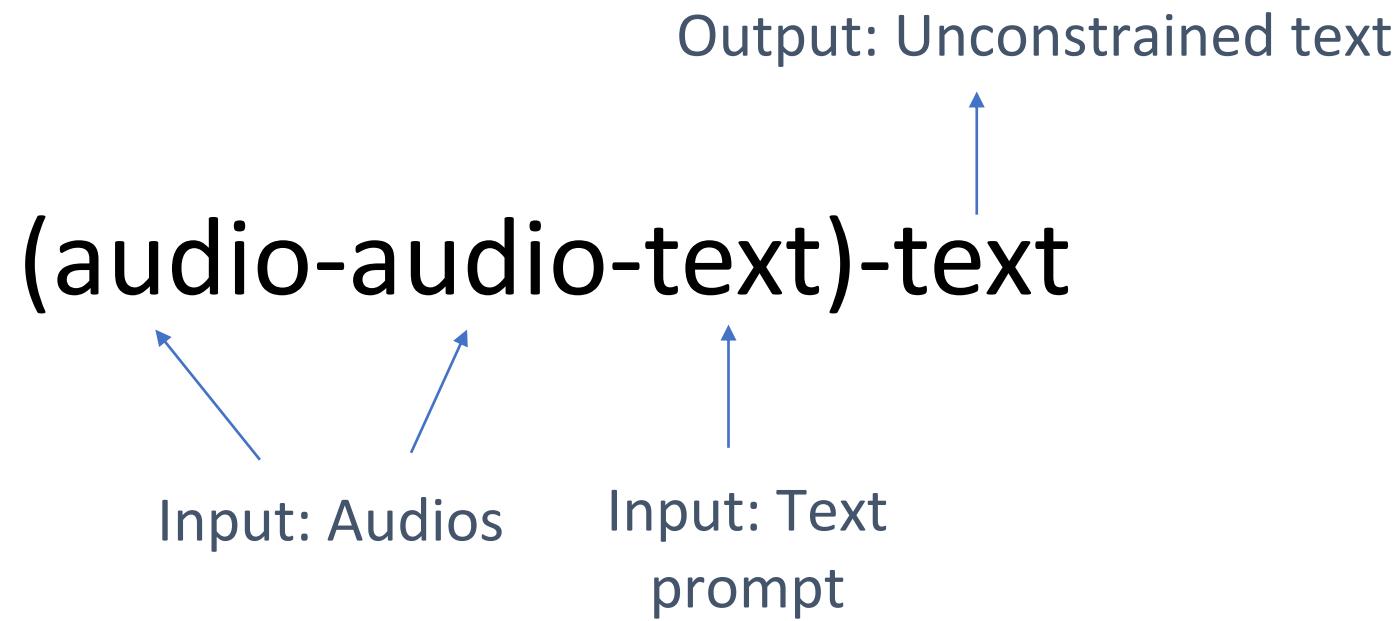
# Mellow architecture



$$\mathcal{L} = - \sum_{i=1}^N \sum_{j=1}^l \log p_\gamma(c_j^i | p_1^i, \dots, p_{2k}^i, c_1^i, \dots, c_{j-1}^i)$$

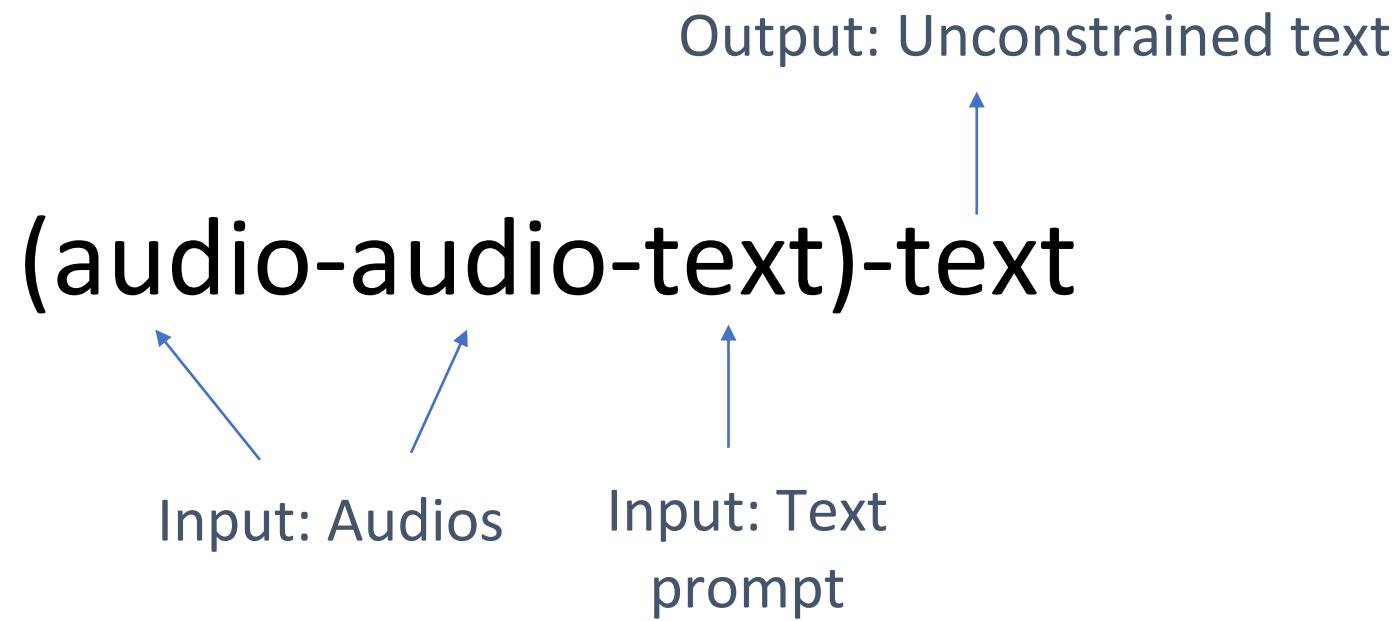
Trained using standard  
captioning loss

# Training data format



**1M instances**

# Training data format



**1M instances - 70% synthetic**

# Results

We benchmark mellow on ~8 understanding and reasoning tasks

- MMAU
- Audio Entailment
- Audio Difference Explanation
- Audio Captioning
- Audio Question Answering
- Audio Understanding
- .....

# Results

We benchmark mellow on ~8 understanding and reasoning tasks

- **MMAU** [OOD and includes both understanding and reasoning]
- Audio Entailment
- Audio Difference Explanation
- Audio Captioning
- Audio Question Answering
- Audio Understanding
- .....

# Results on MMAU

Models	Size	{So, Mu, Sp}	Sound		Music		Speech		Avg	
			Test-mini	Test	Test-mini	Test	Test-mini	Test	Test-mini	Test
<b>Large Audio Language Models (LALMs)</b>										
LTU	7B	✓ ✓ ×	22.52	25.86	09.69	12.83	17.71	16.37	16.89	18.51
LTU AS	7B	✓ ✓ ✓	23.35	24.96	9.10	10.46	20.60	21.30	17.68	18.90
MusiLingo	7B	✗ ✓ ×	23.12	27.76	03.96	06.00	05.88	06.42	10.98	13.39
MuLLaMa	7B	✗ ✓ ×	40.84	44.80	32.63	30.63	22.22	16.56	31.90	30.66
M2UGen	7B	✗ ✓ ×	03.60	03.69	32.93	30.40	06.36	04.53	14.28	12.87
GAMA	7B	✓ ✓ ×	41.44	45.40	32.33	30.83	18.91	19.21	30.90	31.81
GAMA-IT	7B	✓ ✓ ×	43.24	43.23	28.44	28.00	18.91	15.84	30.20	29.02
Qwen-Audio-Chat	8.4B	✓ ✗ ×	<u>55.25</u>	<b>56.73</b>	44.00	40.90	30.03	27.95	43.10	41.86
Qwen2-Audio	8.4B	✓ ✓ ✓	07.50	08.20	05.14	06.16	03.10	04.24	05.24	06.20
Qwen2-Audio-Instruct	8.4B	✓ ✓ ✓	54.95	45.90	<b>50.98</b>	<b>53.26</b>	<u>42.04</u>	<u>45.90</u>	<u>49.20</u>	<u>52.50</u>
SALAMONN	13B	✓ ✓ ✓	41.00	40.30	34.80	33.76	25.50	24.24	33.70	32.77
Gemini Pro v1.5	-	-	<b>56.75</b>	<u>54.46</u>	<u>49.40</u>	<u>48.56</u>	<b>58.55</b>	<b>55.90</b>	<b>54.90</b>	<b>52.97</b>
<b>Small Audio Language Models (SALMs)</b>										
Pengi	323M	✓ ✓ ×	06.10	08.00	02.90	03.05	01.20	01.50	03.40	04.18
Audio Flamingo Chat	2.2B	✓ ✓ ×	<u>23.42</u>	<u>28.26</u>	<u>15.26</u>	<u>18.20</u>	<u>11.41</u>	<u>10.16</u>	<u>16.69</u>	<u>18.87</u>
Mellow	167M	✓ ✗ ×	<b>61.26</b>	<b>64.90</b>	<b>54.19</b>	<b>52.67</b>	<b>29.73</b>	<b>38.77</b>	<b>48.40</b>	<b>52.11</b>

# Results on MMAU

Models	Size	{So, Mu, Sp}	Sound		Music		Speech		Avg	
			Test-mini	Test	Test-mini	Test	Test-mini	Test	Test-mini	Test
<b>Large Audio Language Models (LALMs)</b>										
LTU	7B	✓ ✓ ×	22.52	25.86	09.69	12.83	17.71	16.37	16.89	18.51
LTU AS	7B	✓ ✓ ✓	23.35	24.96	9.10	10.46	20.60	21.30	17.68	18.90
MusiLingo	7B	✗ ✓ ×	23.12	27.76	03.96	06.00	05.88	06.42	10.98	13.39
MuLLaMa	7B	✗ ✓ ×	40.84	44.80	32.63	30.63	22.22	16.56	31.90	30.66
M2UGen	7B	✗ ✓ ×	03.60	03.69	32.93	30.40	06.36	04.53	14.28	12.87
GAMA	7B	✓ ✓ ×	41.44	45.40	32.33	30.83	18.91	19.21	30.90	31.81
GAMA-IT	7B	✓ ✓ ×	43.24	43.23	28.44	28.00	18.91	15.84	30.20	29.02
Qwen-Audio-Chat	8.4B	✓ ✗ ✗	<u>55.25</u>	<b>56.73</b>	44.00	40.90	30.03	27.95	43.10	41.86
Qwen2-Audio	8.4B	✓ ✓ ✓	07.50	08.20	05.14	06.16	03.10	04.24	05.24	06.20
Qwen2-Audio-Instruct	8.4B	✓ ✓ ✓	54.95	45.90	<b>50.98</b>	<b>53.26</b>	<u>42.04</u>	<u>45.90</u>	<u>49.20</u>	<u>52.50</u>
SALAMONN	13B	✓ ✓ ✓	41.00	40.30	34.80	33.76	25.50	24.24	33.70	32.77
Gemini Pro v1.5	-	-	<b>56.75</b>	<u>54.46</u>	<u>49.40</u>	<u>48.56</u>	<b>58.55</b>	<b>55.90</b>	<b>54.90</b>	<b>52.97</b>
<b>Small Audio Language Models (SALMs)</b>										
Pengi	323M	✓ ✓ ×	06.10	08.00	02.90	03.05	01.20	01.50	03.40	04.18
Audio Flamingo Chat	2.2B	✓ ✓ ×	<u>23.42</u>	<u>28.26</u>	<u>15.26</u>	<u>18.20</u>	<u>11.41</u>	<u>10.16</u>	<u>16.69</u>	<u>18.87</u>
Mellow	167M	✓ ✗ ✗	<b>61.26</b>	<b>64.90</b>	<b>54.19</b>	<b>52.67</b>	<b>29.73</b>	<b>38.77</b>	<b>48.40</b>	<b>52.11</b>

# Results on MMAU

Models	Size	{So, Mu, Sp}	Sound		Music		Speech		Avg	
			Test-mini	Test	Test-mini	Test	Test-mini	Test	Test-mini	Test
<b>Large Audio Language Models (LALMs)</b>										
LTU	7B	✓ ✓ ×	22.52	25.86	09.69	12.83	17.71	16.37	16.89	18.51
LTU AS	7B	✓ ✓ ✓	23.35	24.96	9.10	10.46	20.60	21.30	17.68	18.90
MusiLingo	7B	✗ ✓ ×	23.12	27.76	03.96	06.00	05.88	06.42	10.98	13.39
MuLLaMa	7B	✗ ✓ ×	40.84	44.80	32.63	30.63	22.22	16.56	31.90	30.66
M2UGen	7B	✗ ✓ ×	03.60	03.69	32.93	30.40	06.36	04.53	14.28	12.87
GAMA	7B	✓ ✓ ×	41.44	45.40	32.33	30.83	18.91	19.21	30.90	31.81
GAMA-IT	7B	✓ ✓ ×	43.24	43.23	28.44	28.00	18.91	15.84	30.20	29.02
Qwen-Audio-Chat	8.4B	✓ ✗ ✗	<u>55.25</u>	<b>56.73</b>	44.00	40.90	30.03	27.95	43.10	41.86
Owen2-Audio	8.4B	✓ ✓ ✓	07.50	08.20	05.14	06.16	03.10	04.24	05.24	06.20
Qwen2-Audio-Instruct	8.4B	✓ ✓ ✓	54.95	45.90	<b>50.98</b>	<b>53.26</b>	<u>42.04</u>	<u>45.90</u>	<u>49.20</u>	<u>52.50</u>
SALAMONN	13B	✓ ✓ ✓	41.00	40.30	34.80	33.76	25.50	24.24	33.70	32.77
Gemini Pro v1.5	-	-	<b>56.75</b>	<u>54.46</u>	<u>49.40</u>	<u>48.56</u>	<b>58.55</b>	<b>55.90</b>	<b>54.90</b>	<b>52.97</b>
<b>Small Audio Language Models (SALMs)</b>										
Pengi	323M	✓ ✓ ×	06.10	08.00	02.90	03.05	01.20	01.50	03.40	04.18
Audio Flamingo Chat	2.2B	✓ ✓ ✗	23.42	28.26	15.26	18.20	11.41	10.16	16.69	18.87
Mellow	167M	✓ ✗ ✗	<b>61.26</b>	<b>64.90</b>	<b>54.19</b>	<b>52.67</b>	<b>29.73</b>	<b>38.77</b>	<b>48.40</b>	<b>52.11</b>

# Ablations

Models	Size	Audio Caption		B-AQA	Audio Entailment		Audio Difference		MMAU (test-mini)				
		AC (SPICE)	CL (SPICE)	ClothoAQA (ACC)	CLE (ACC)	ACE (ACC)	CLD-3 (SPICE)	ACD-3 (SPICE)	Sound (ACC)	Music (ACC)	Speech (ACC)	Avg. (ACC)	
<i>The SLM is GPT2 and frozen and the projection layer is changed</i>													
Linear	156M	4.53	5.98	53.23	30.79	30.15	5.21	8.39	16.80	31.25	30.10	26.05	
Non-linear [2]	157M	4.79	6.10	55.18	32.12	29.56	5.56	8.62	17.42	33.23	27.03	25.89	
Transformer [14]	195M	10.26	7.89	62.35	43.89	42.10	11.88	12.50	28.65	35.23	26.68	30.19	
<i>The SLM is GPT2 and finetuned and the projection layer is changed</i>													
Linear	156M	9.87	7.10	70.90	93.45	92.90	13.65	14.21	47.64	45.39	27.00	40.01	
Non-linear [2]	157M	10.51	7.15	71.25	93.40	93.27	13.77	14.01	48.05	48.50	27.33	41.29	
Transformer [14]	195M	10.77	7.23	70.89	92.32	93.65	13.45	15.21	47.89	49.10	27.10	41.36	
<i>The SLM is frozen, projection layer is non-linear and the SLM is changed</i>													
GPT2 frozen	157M	4.79	6.10	55.18	32.12	29.56	5.56	8.62	17.42	33.23	27.03	25.89	
SmolLM2 frozen	167M	5.26	8.58	45.70	33.40	31.66	11.43	13.62	35.14	30.54	20.12	28.60	
<i>The SLM is finetuned, projection layer is non-linear and the SLM is changed</i>													
GPT2 finetune	157M	10.51	7.15	71.25	93.40	93.27	13.77	14.01	48.05	48.50	27.33	41.29	
SmolLM2 finetune	167M	18.60	9.83	71.65	92.00	90.85	17.33	18.68	59.46	50.60	28.82	46.29	
<i>Different finetuning methods. The SLM is SmolLM2, projection layer is non-linear</i>													
Prefix-tuning	167M	5.26	8.58	45.70	33.40	31.66	11.43	13.62	35.14	30.54	20.12	28.60	
LoRA (8, 16)	167M	18.53	9.25	64.64	79.33	84.15	14.23	14.98	45.95	42.51	28.83	39.10	
LoRA (256, 512)	181M	19.01	10.59	65.54	86.66	89.87	15.36	16.51	50.75	49.10	33.33	44.40	
Finetuning	167M	18.60	9.83	71.65	92.00	90.85	17.33	18.68	59.46	50.60	28.82	46.29	
<i>The SLM is SmolLM2 and finetuned, projection layer is non-linear and the audio encoder is changed</i>													
CNN14	219M	15.97	7.91	65.82	91.06	92.39	16.27	16.95	54.05	47.60	28.23	43.30	
HTSAT	167M	18.60	9.83	71.65	92.00	90.85	17.33	18.68	59.46	50.60	28.82	46.29	
<i>The SLM is SmolLM2 and finetuned, projection layer is non-linear, and training data is changed</i>													
Type 1	167M	18.60	9.83	71.65	92.00	90.85	17.33	18.68	59.46	50.60	28.82	46.29	
Type 2	167M	16.47	8.23	71.05	92.50	93.20	16.98	18.09	59.45	42.81	37.84	46.70	
Type 3	167M	17.43	9.88	66.83	91.87	94.25	17.37	18.67	61.56	45.21	32.43	46.40	
Type 4	167M	17.79	9.38	71.39	91.16	89.66	17.21	18.54	61.26	54.19	29.73	48.40	
<i>The SLM is SmolLM2 and finetuned, projection layer is non-linear, and WavCaps is added to training</i>													
Type 1	167M	18.60	9.83	71.65	92.00	90.85	17.33	18.68	59.46	50.60	28.82	46.29	
+ WavCaps	167M	14.83	9.66	71.32	92.47	92.69	17.92	19.13	59.16	60.48	23.72	47.80	
<b>Mellow</b>	167M	17.79	9.38	71.39	91.16	89.66	17.21	18.54	61.26	54.19	29.73	48.40	

# Ablations

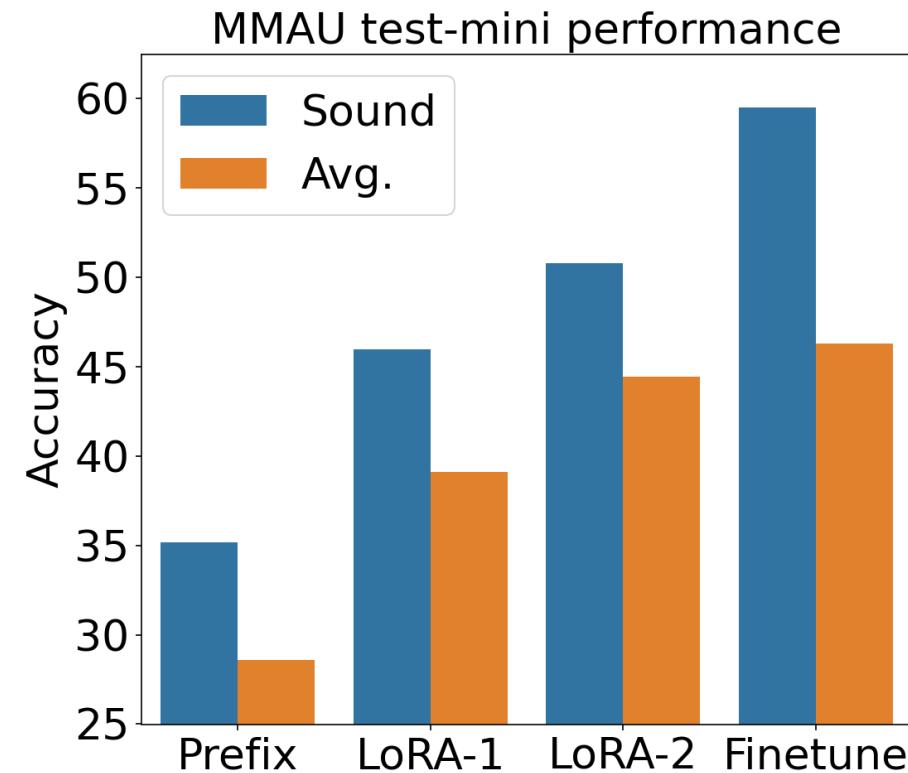
Models	Size	Audio Caption		B-AQA	Audio Entailment		Audio Difference		MMAU (test-mini)			
		AC (SPICE)	CL (SPICE)	ClothoAQA (ACC)	CLE (ACC)	ACE (ACC)	CLD-3 (SPICE)	ACD-3 (SPICE)	Sound (ACC)	Music (ACC)	Speech (ACC)	Avg. (ACC)
<i>The SLM is GPT2 and frozen and the projection layer is changed</i>												
Linear	156M	4.53	5.98	53.23	30.79	30.15	5.21	8.39	16.80	31.25	30.10	26.05
Non-linear [2]	157M	4.79	6.10	55.18	32.12	29.56	5.56	8.62	17.42	33.23	27.03	25.89
Transformer [14]	195M	10.26	7.89	62.35	43.89	42.10	11.88	12.50	28.65	35.23	26.68	30.19
<i>The SLM is GPT2 and finetuned and the projection layer is changed</i>												
Linear	156M	9.87	7.10	70.90	93.45	92.90	13.65	14.21	47.64	45.39	27.00	40.01
Non-linear [2]	157M	10.51	7.15	71.25	93.40	93.27	13.77	14.01	48.05	48.50	27.33	41.29
Transformer [14]	195M	10.77	7.23	70.89	92.32	93.65	13.45	15.21	47.89	49.10	27.10	41.36
<i>The SLM is frozen, projection layer is non-linear and the SLM is changed</i>												
GPT2 frozen	157M	4.79	6.10	55.18	32.12	29.56	5.56	8.62	17.42	33.23	27.03	25.89
SmolLM2 frozen	167M	5.26	8.58	45.70	33.40	31.66	11.43	13.62	35.14	30.54	20.12	28.60
<i>The SLM is finetuned, projection layer is non-linear and the SLM is changed</i>												
GPT2 finetune	157M	10.51	7.15	71.25	93.40	93.27	13.77	14.01	48.05	48.50	27.33	41.29
SmolLM2 finetune	167M	18.60	9.83	71.65	92.00	90.85	17.33	18.68	59.46	50.60	28.82	46.29
<i>Different finetuning methods. The SLM is SmolLM2, projection layer is non-linear</i>												
Prefix-tuning	167M	5.26	8.58	45.70	33.40	31.66	11.43	13.62	35.14	30.54	20.12	28.60
LoRA (8, 16)	167M	18.53	9.25	64.64	79.33	84.15	14.23	14.98	45.95	42.51	28.83	39.10
LoRA (256, 512)	181M	19.01	10.59	65.54	86.66	89.87	15.36	16.51	50.75	49.10	33.33	44.40
Finetuning	167M	18.60	9.83	71.65	92.00	90.85	17.33	18.68	59.46	50.60	28.82	46.29
<i>The SLM is SmolLM2 and finetuned, projection layer is non-linear and the audio encoder is changed</i>												
CNN14	219M	15.97	7.91	65.82	91.06	92.39	16.27	16.95	54.05	47.60	28.23	43.30
HTSAT	167M	18.60	9.83	71.65	92.00	90.85	17.33	18.68	59.46	50.60	28.82	46.29
<i>The SLM is SmolLM2 and finetuned, projection layer is non-linear, and training data is changed</i>												
Type 1	167M	18.60	9.83	71.65	92.00	90.85	17.33	18.68	59.46	50.60	28.82	46.29
Type 2	167M	16.47	8.23	71.05	92.50	93.20	16.98	18.09	59.45	42.81	37.84	46.70
Type 3	167M	17.43	9.88	66.83	91.87	94.25	17.37	18.67	61.56	45.21	32.43	46.40
Type 4	167M	17.79	9.38	71.39	91.16	89.66	17.21	18.54	61.26	54.19	29.73	48.40
<i>The SLM is SmolLM2 and finetuned, projection layer is non-linear, and WavCaps is added to training</i>												
Type 1	167M	18.60	9.83	71.65	92.00	90.85	17.33	18.68	59.46	50.60	28.82	46.29
+ WavCaps	167M	14.83	9.66	71.32	92.47	92.69	17.92	19.13	59.16	60.48	23.72	47.80
<b>Mellow</b>	167M	17.79	9.38	71.39	91.16	89.66	17.21	18.54	61.26	54.19	29.73	48.40

Pengi, random performance

Mellow, SoTA performance

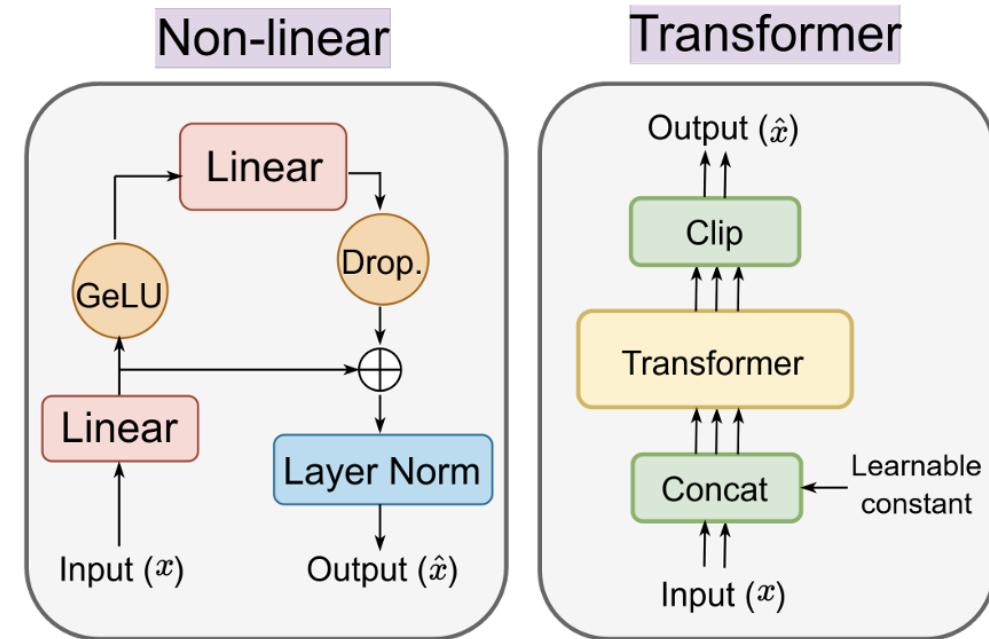
# Ablation findings

1. Fine-tuning outperforms prefix-tuning



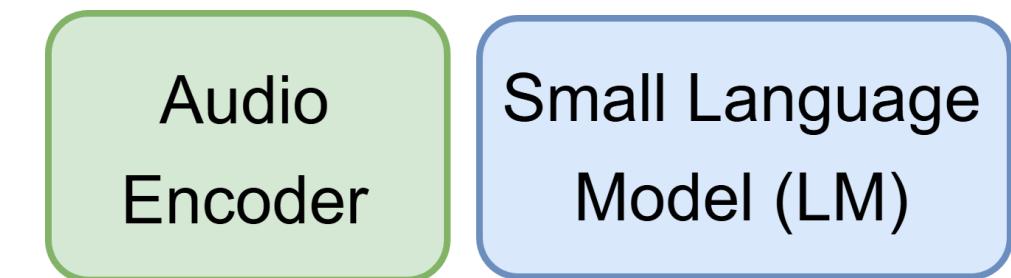
# Ablation findings

1. Fine-tuning outperforms prefix-tuning
2. Simple non-linear projections reduces overfitting for LoRA and finetuning leading to better performance



# Ablation findings

1. Fine-tuning outperforms prefix-tuning
2. Simple non-linear projections reduces overfitting for LoRA and finetuning leading to better performance
3. Unimodal pretraining influences multimodal capabilities



# Ablation findings

1. Fine-tuning outperforms prefix-tuning
2. Simple non-linear projections reduces overfitting for LoRA and finetuning leading to better performance
3. Unimodal pretraining influences multimodal capabilities
4. Reasoning-focused synthetic data addition enhances performance

## *ReasonQA examples*

**Question:** What is the sound event present in the clip?

**Answer:** The sound event present in the clip is a gunshot, accompanied by background conversations and ambient noise. The gunshot is a sudden, sharp, and loud sound, followed by a decaying reverberation, indicating a sudden release of energy. The background conversations and ambient noise provide a sense of context, suggesting a chaotic or noisy environment.

**Question:** What is the dominant acoustic property of the gunshot?

**Answer:** The dominant acoustic property of the gunshot is its high-frequency content, particularly in the range of 5 kHz to 20 kHz. This is because the sudden release of energy from the gunpowder creates a high-pressure wave that propagates through the air, producing a sharp, crisp sound with a high-frequency emphasis.

**Question:** How does the gunshot sound different from the background conversations?

**Answer:** The gunshot sounds different from the background conversations in terms of its spectral characteristics, duration, and intensity. The gunshot has a much higher intensity, with a peak amplitude that is significantly higher than the background conversations. The gunshot also has a much shorter duration, lasting around 0.1-0.2 seconds, whereas the background conversations are more sustained and have a longer duration.

**Question:** What is the psychological impact of hearing this sound?

**Answer:** Hearing this sound can have a significant psychological impact, particularly due to its association with violence, danger, or sudden, intense events. The sudden and intense nature of the gunshot can trigger a fight-or-flight response, releasing stress hormones such as adrenaline and cortisol. The background conversations may also contribute to a sense of chaos or confusion, making it difficult to process the information and respond appropriately.

**Question:** What is this sound most similar to?

**Answer:** This sound is most similar to other sudden, intense sounds, such as firecrackers, car crashes, or explosions. The sharp, high-frequency content and sudden release of energy are common characteristics of these types of sounds, making them easily recognizable and evocative of strong emotions and reactions.

# Talk outline

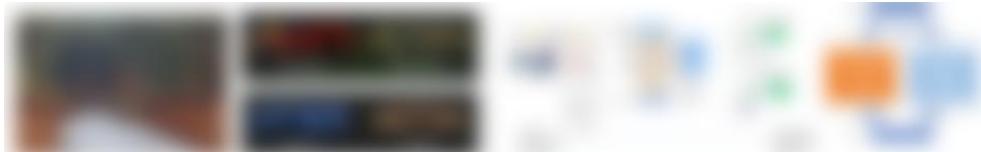
- Effective learning and pretraining methods  
CLAP (ICASSP 23, 24), Pengi (NeurIPS 23), Mellow (under review)



- Methods or benchmarks to analyze new abilities  
Audio Entail (AAAI 25), Audio Difference (ICLR 25), PAM (INTERSPEECH 24)



- Efficient adaptation methods to enable customization  
Test-time (ICASSP 24), Few-shot (INTERSPEECH 25), Missing modality (ICASSP 25)



# Talk outline

- Effective learning and pretraining methods  
CLAP (ICASSP 23, 24), Pengi (NeurIPS 23), Mellow (under review)



- **Methods or benchmarks to analyze new abilities**

Audio Entail (AAAI 25), Audio Difference (ICLR 25), PAM (INTERSPEECH 24)

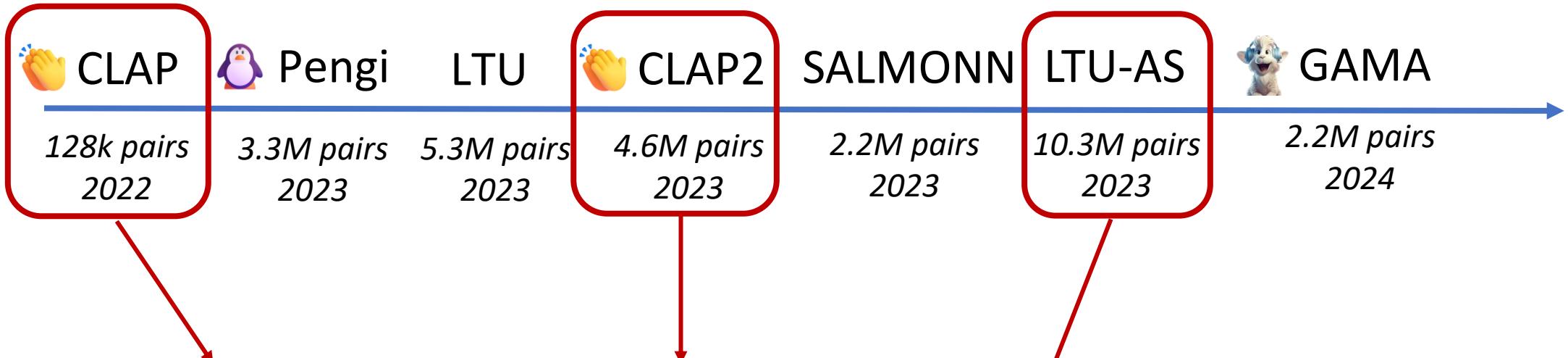


- Efficient adaptation methods to enable customization

Test-time (ICASSP 24), Few-shot (INTERSPEECH 25), Missing modality (ICASSP 25)

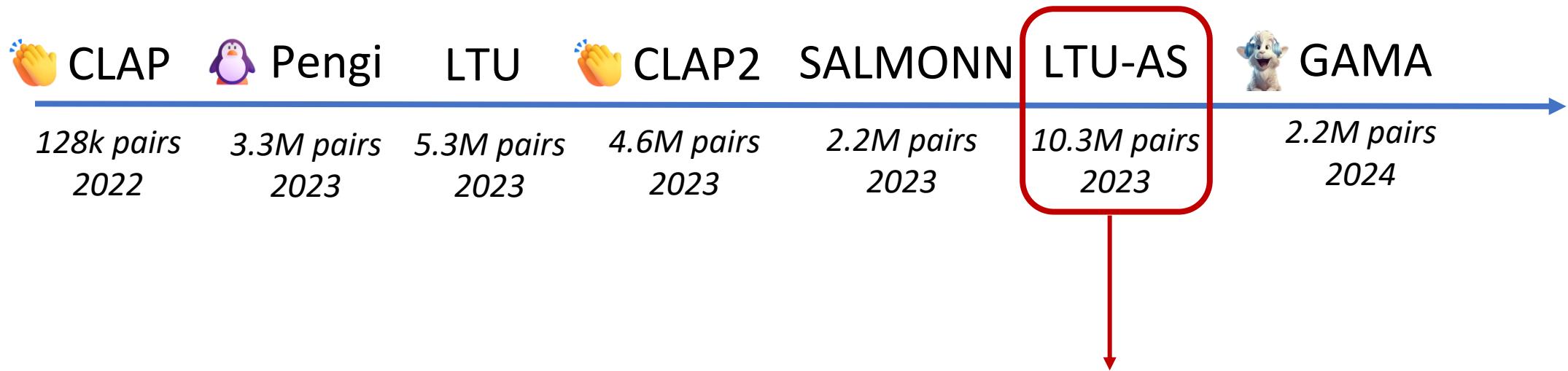


# Scaling data and compute



Audio foundation models are trained on **increasingly larger web-scale data**

# Scaling data and compute



With scale, the model starts to **encapsulate world-knowledge, exhibit emergent abilities and show primitive reasoning**

# Consider this conversation with audio assistant

Hi, I am an Audio Assistant, how may I help you?



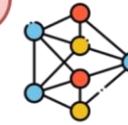
# Consider this conversation with audio assistant



I recorded my car engine's sound on my iPhone. Any chance you could tell me if it sounds like something serious?



Hi, I am an Audio Assistant, how may I help you?



# Consider this conversation with audio assistant

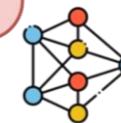


I recorded my car engine's sound on my iPhone. Any chance you could tell me if it sounds like something serious?

Hi, I am an Audio Assistant, how may I help you?



It seems your engine is experiencing a knocking sound, which might indicate detonation. It is likely caused by fuel issues or a misfiring cylinder.



# Consider this conversation with audio assistant



I recorded my car engine's sound on my iPhone. Any chance you could tell me if it sounds like something serious?



What should I do?

Hi, I am an Audio Assistant, how may I help you?



It seems your engine is experiencing a knocking sound, which might indicate detonation. It is likely caused by fuel issues or a misfiring cylinder.



# Consider this conversation with audio assistant

The diagram illustrates a conversation between a user and an audio assistant. It features three blue speech bubbles for the user and three pink speech bubbles for the audio assistant. The user's first message includes a sound波 icon. The audio assistant's responses include icons of neural network graphs.

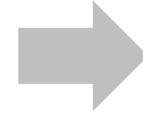
**User (Blue Bubble):**

- I recorded my car engine's sound on my iPhone. Any chance you could tell me if it sounds like something serious?
- What should I do?

**Audio Assistant (Pink Bubble):**

- Hi, I am an Audio Assistant, how may I help you?
- It seems your engine is experiencing a knocking sound, which might indicate detonation. It is likely caused by fuel issues or a misfiring cylinder.
- I recommend having a mechanic check your fuel injectors and spark plugs. If left unchecked, this issue could damage the engine.

# The audio assistant shows primitive reasoning abilities in the conversation



**Perception and understanding**  
engine knocking sound

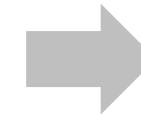
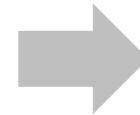


**World knowledge**  
car sounds, engine issues,  
misfiring cylinder ..



**Logical deduction**  
Indicate detonation. It is likely caused by fuel issues or a misfiring cylinder

# The audio assistant shows primitive reasoning abilities in the conversation



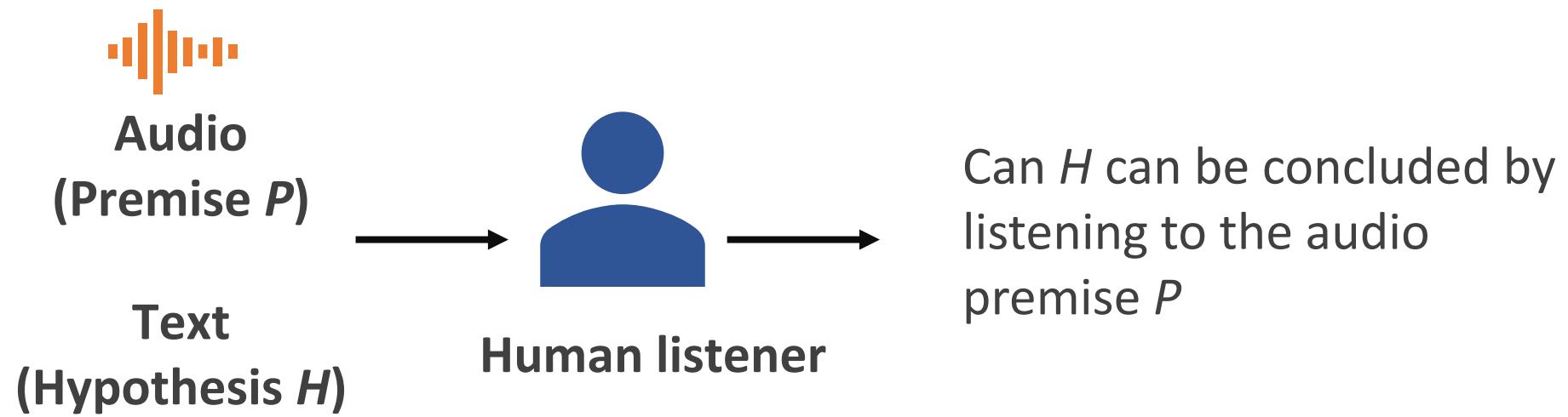
**Perception and understanding**  
engine knocking sound

**World knowledge**  
car sounds, engine issues,  
misfiring cylinder ..

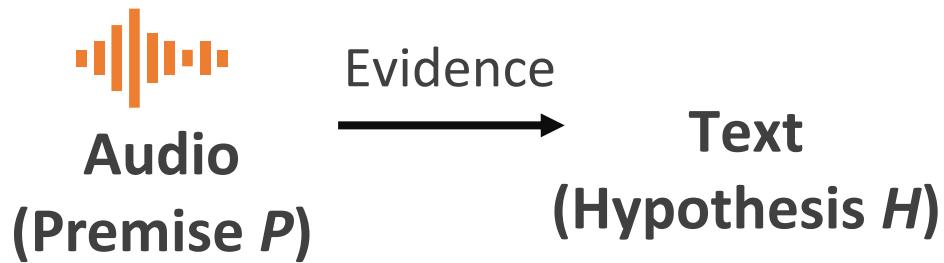
**Logical deduction**  
Indicate detonation. It is likely caused by fuel issues or a misfiring cylinder

Benchmarking logical reasoning abilities is necessary to estimate performance in real-world scenarios

# To evaluate auditory deductive reasoning for humans

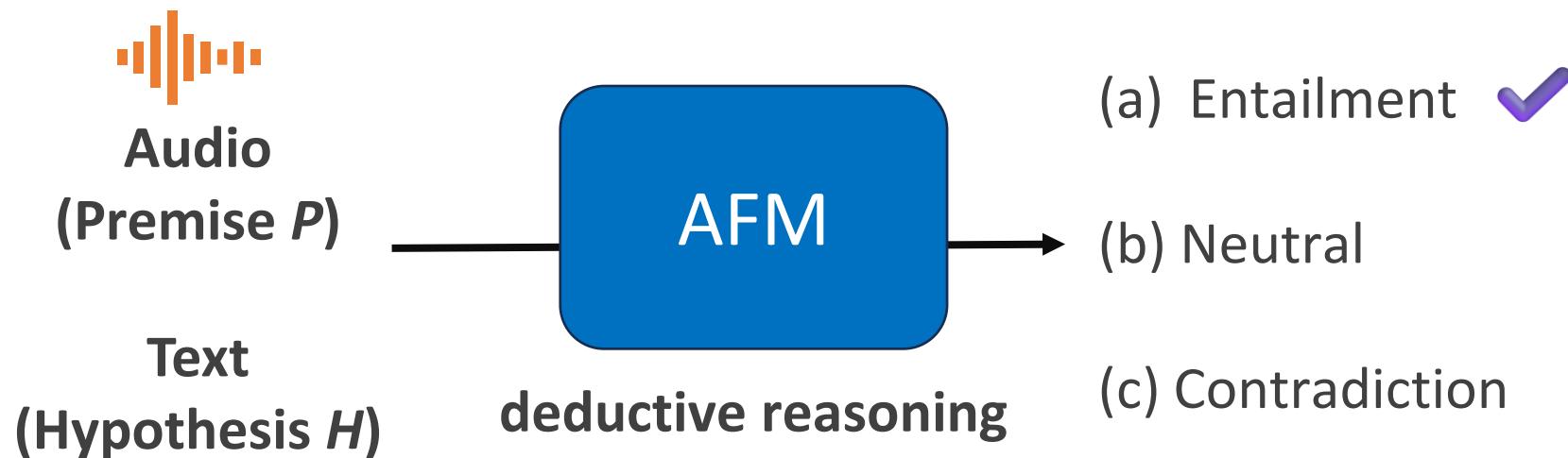


# Three possible scenarios

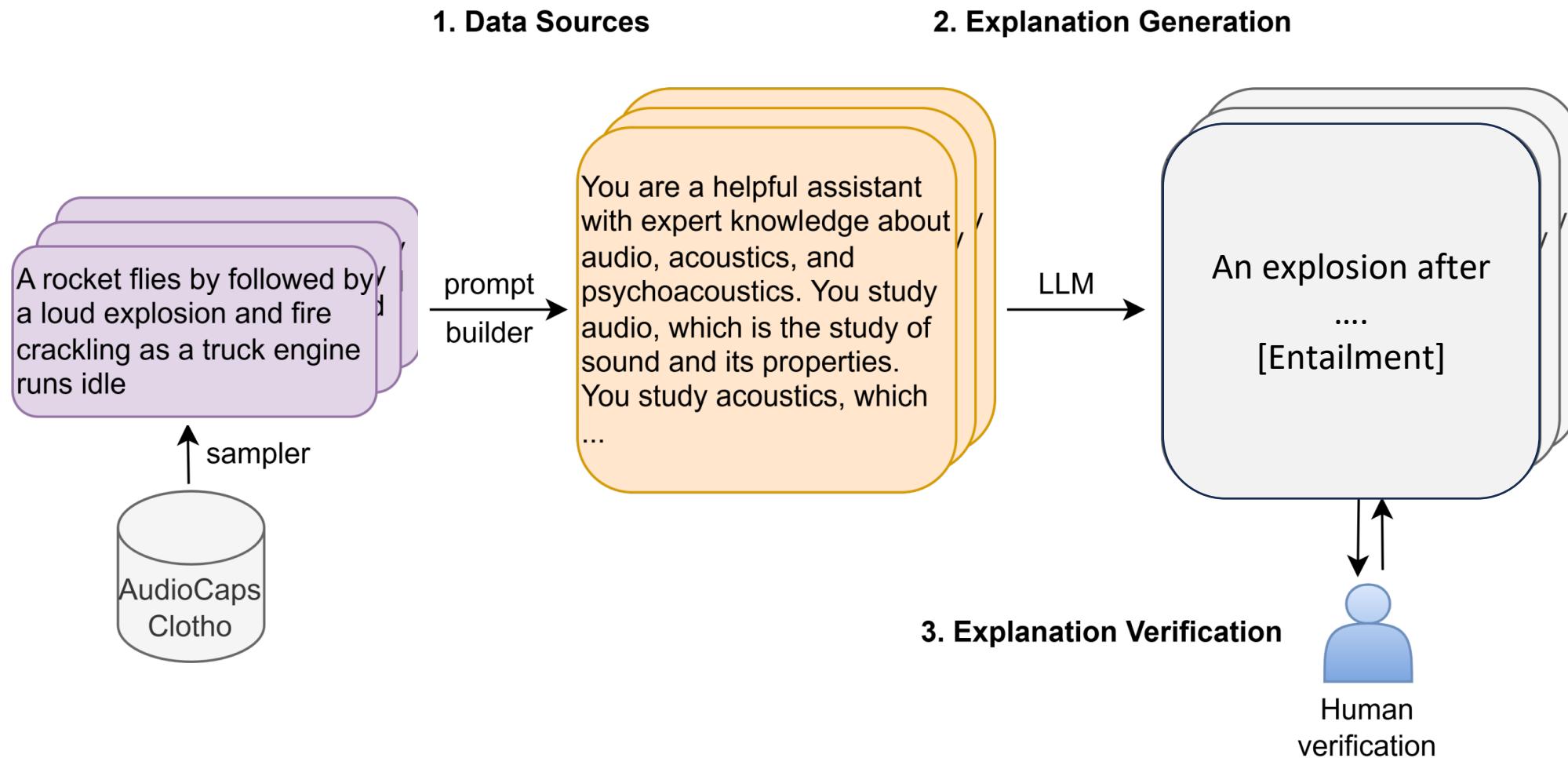


1.  $P$  contains sufficient evidence to affirm the truth of  $H$
2.  $P$  does not have enough information to either confirm or deny  $H$
3.  $P$  have substantial evidence to deduce that  $H$  is false.

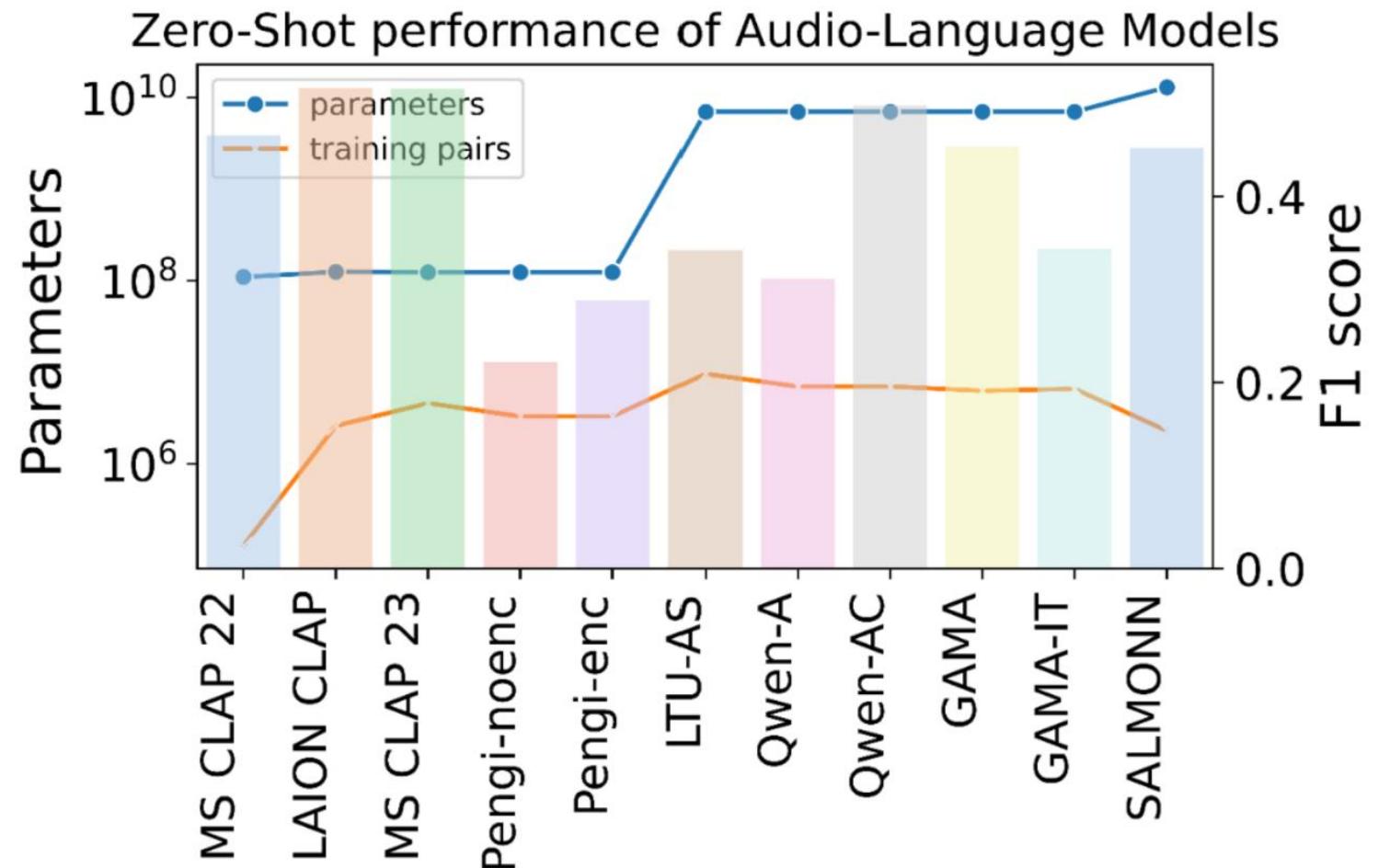
# Introduce Audio Entailment task



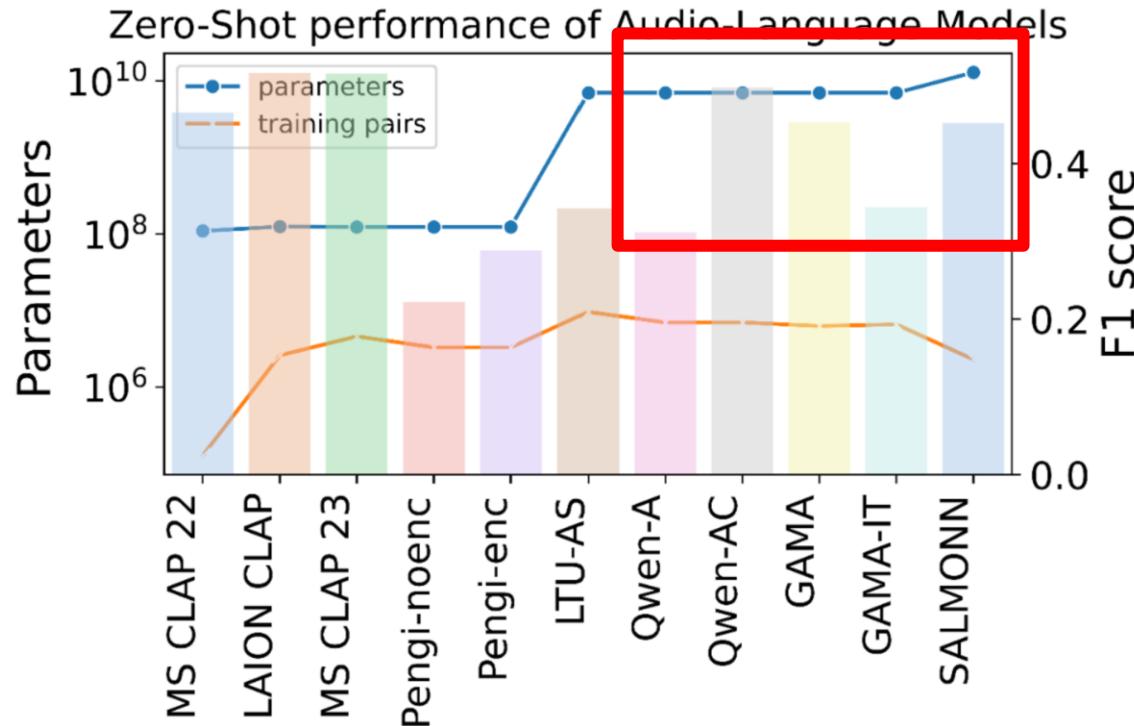
# Audio Entailment task dataset



# Benchmarking AFM on Audio Entailment task



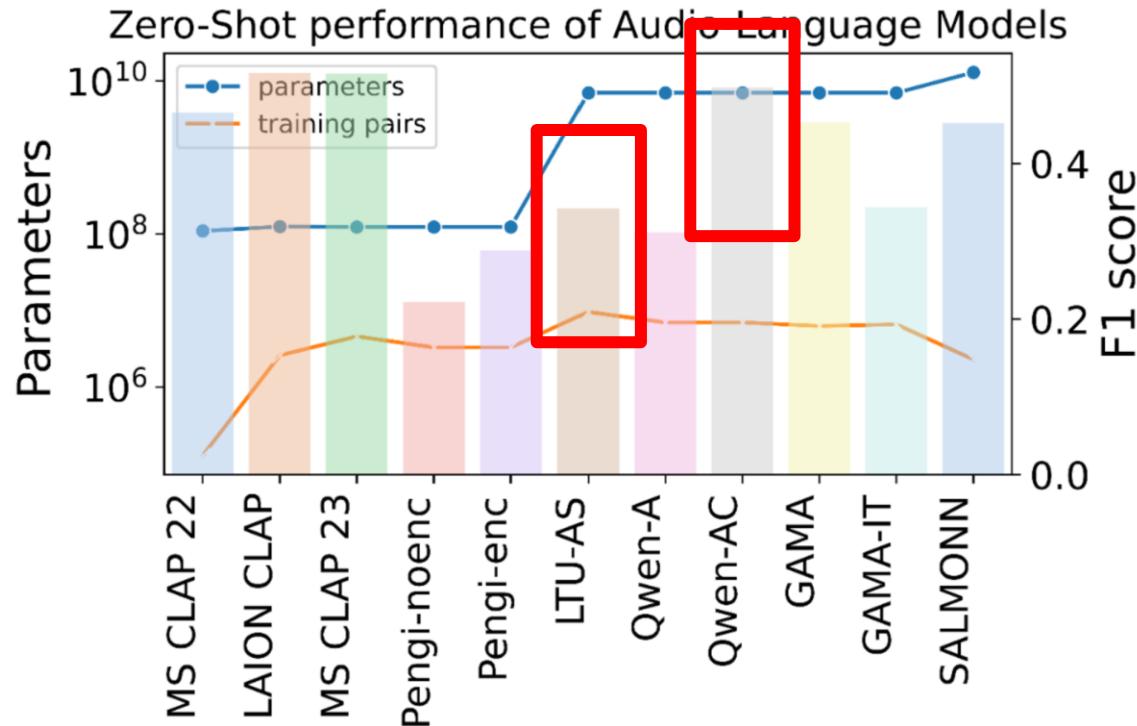
# Findings from benchmarking



(1) Larger language models improve deductive reasoning but are challenging to ground in audio (hallucinate)

Changing stopwords like “it” to “the” in the prompts of SALMONN and GAMA, leads to them changing the deductions

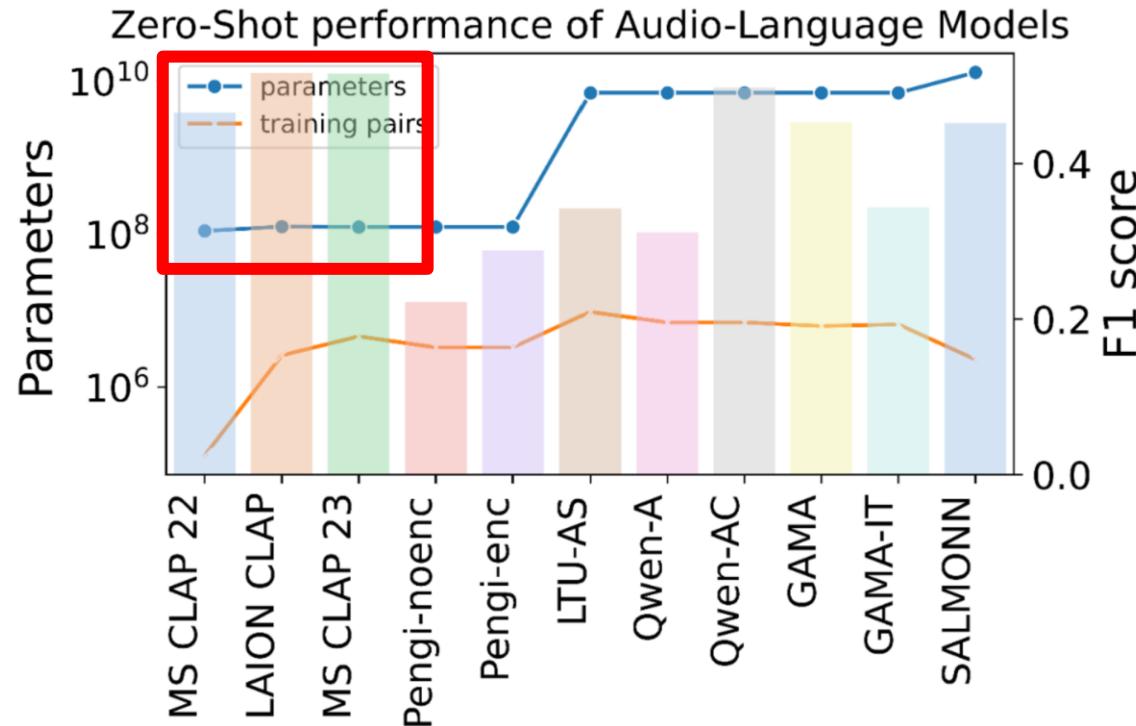
# Findings from benchmarking



(2) Training AFMs to predict uncertainty improves their ability to detect plausible scenario

GAMA and LTU, trained with 6.5% "I don't know" or "cannot answer due to insufficient information" data, better predict when audio lacks sufficient evidence to confirm or deny a hypothesis, but only if prompts align with training data.

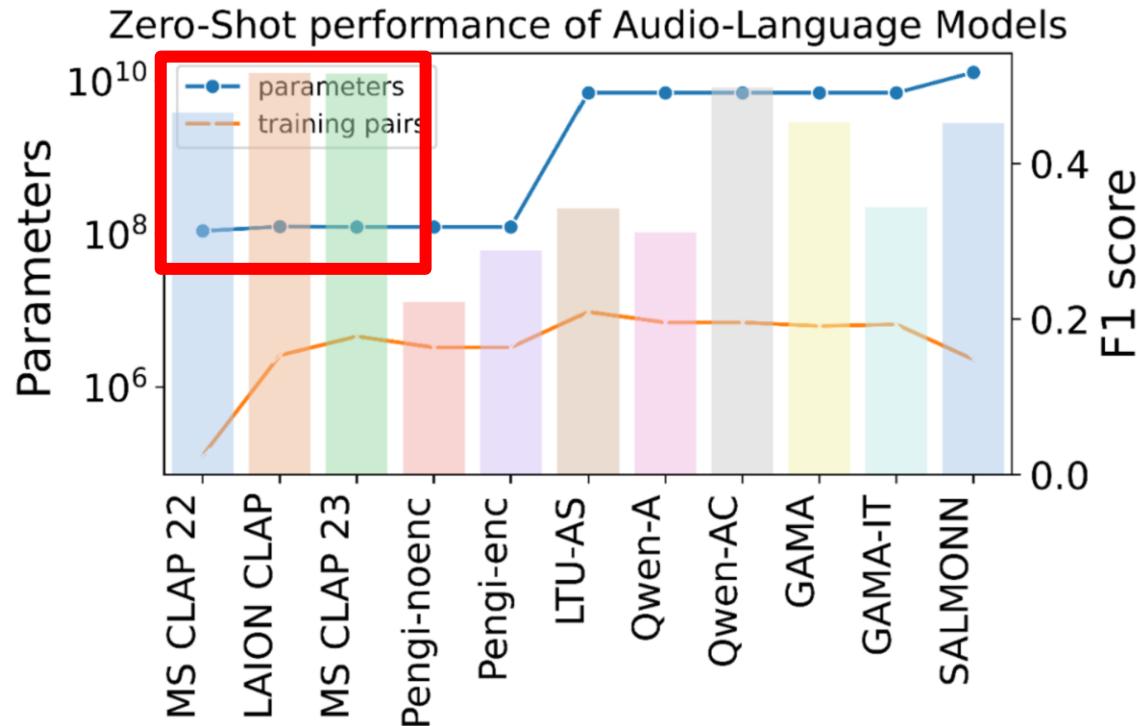
# Findings from benchmarking



(3) Contrastive models are competitive on the task of deductive reasoning\*

Despite nonoverlapping, linearly increasing thresholds, F1 scores are around 50%, showing the CLAP similarity score changes linearly with hypothesis-audio premise closeness.

# Findings from benchmarking



(3) Contrastive models are competitive on the task of deductive reasoning\*

Despite nonoverlapping, linearly increasing thresholds, F1 scores are around 50%, showing the CLAP similarity score changes linearly with hypothesis-audio premise closeness.

How do we improve deductive reasoning of AFMs at test-time?

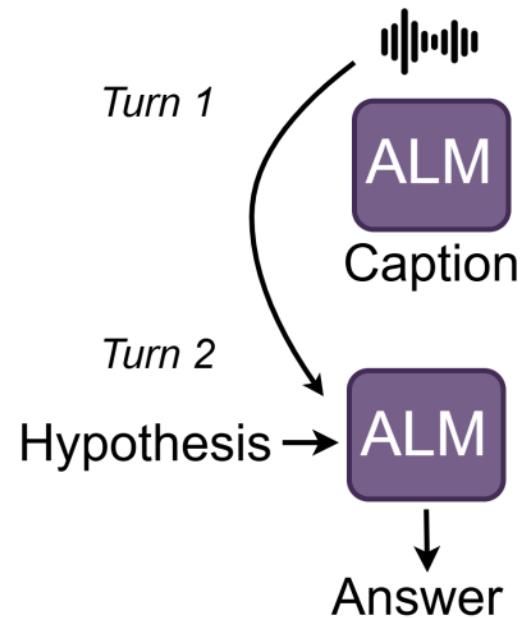
# Reduce perception gap

Caption the audio before performing deductive reasoning

# Reduce perception gap

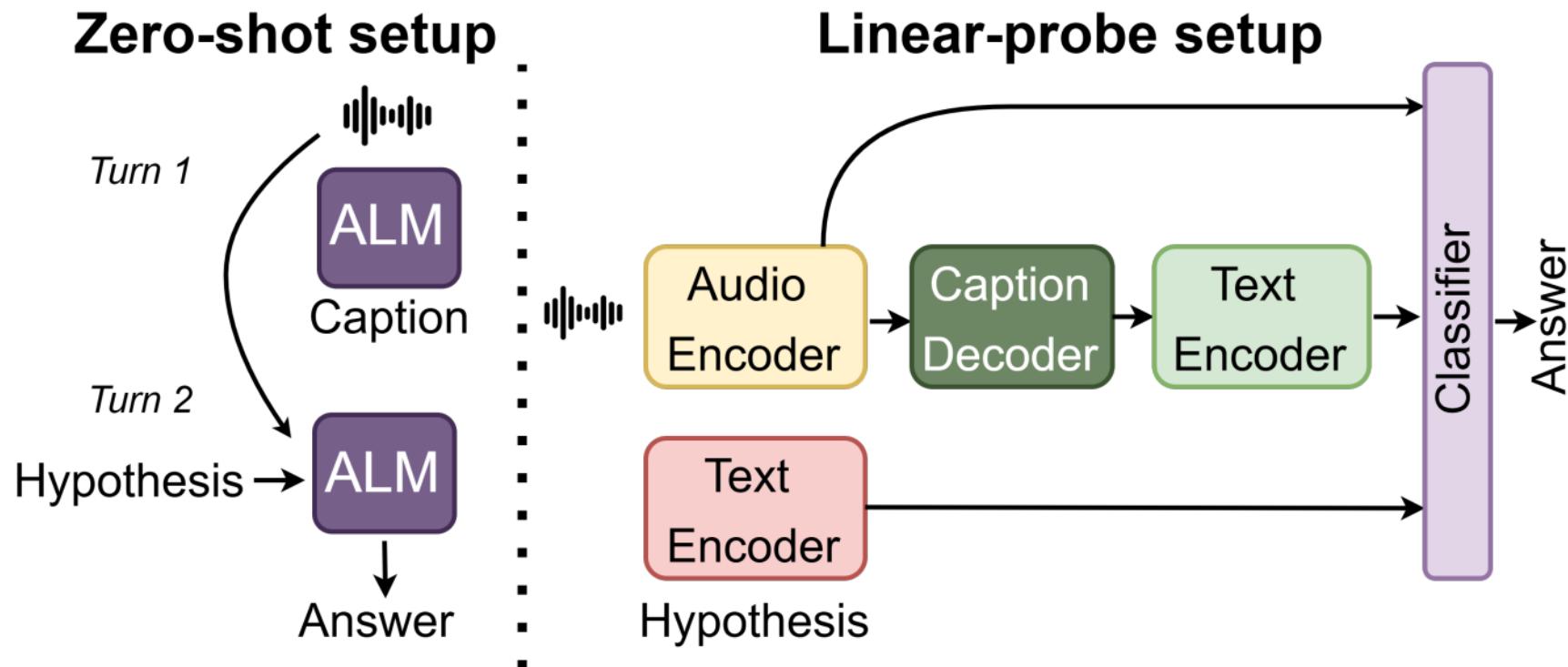
Caption the audio before performing deductive reasoning

## Zero-shot setup



# Reduce perception gap

Caption the audio before performing deductive reasoning



# Reduce perception gap

Caption the audio before performing deductive reasoning

Model	Method	ACC	P	R	F1
Qwen-AC	base	0.5442	0.5604	0.5442	0.4975
Qwen-AC	cap	<b>0.6083</b>	<b>0.5964</b>	<b>0.6083</b>	<b>0.5601</b>
CLAP 23	concat	0.8329	0.8361	0.8329	0.8336
CLAP 23	cap	<b>0.8640</b>	<b>0.8671</b>	<b>0.8640</b>	<b>0.8647</b>

**Improves performance by 6% for Zero-shot and 3% for the Linear-probe setup**

# Humans have the ability to do comparative reasoning over audio

1



# Humans have the ability to do comparative reasoning over audio

1



---

dog barking, cat  
meowing  
park

low intensity

bark after meow

calming

# Humans have the ability to do comparative reasoning over audio

1



---

dog barking, cat  
meowing  
park

low intensity

bark after meow

calming

2



# Humans have the ability to do comparative reasoning over audio

1



dog barking, cat  
meowing  
park

low intensity

bark after meow

calming

2



dog barking, door  
closing  
indoor, kitchen

loud and high  
pitched

bark before  
door closing

intimidating

# Humans have the ability to do comparative reasoning over audio

1



dog barking, cat  
meowing  
park

low intensity

bark after meow

calming

2



dog barking, door  
closing  
indoor, kitchen

loud and high  
pitched

bark before  
door closing

intimidating

3

## Deduction

The two audio clips do not belong to the same scene because ...

# Humans have the ability to do comparative reasoning over audio

1



dog barking, cat  
meowing  
park

low intensity

bark after meow  
calming

2



dog barking, door  
closing  
indoor, kitchen

loud and high  
pitched  
bark before  
door closing  
intimidating

3

## Deduction

The two audio clips do not belong to the same scene because ...

**How to evaluate and subsequently improve this ability for AFMs?**

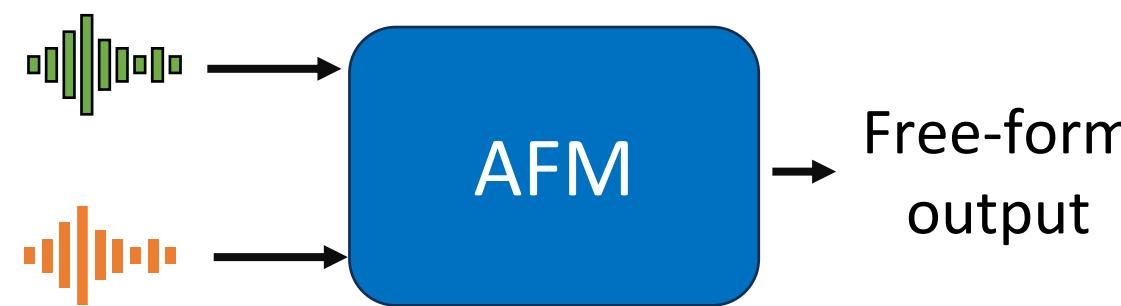
# Introduce Audio Difference Explanation task

Ask the model to explain the difference between two audios

ADIFF: Explaining audio difference using  
natural language,  
<https://arxiv.org/abs/2502.04476> ICLR 2025  
Spotlight

# The Audio Difference Explanation task setup

Ask the model to explain the difference between two audios



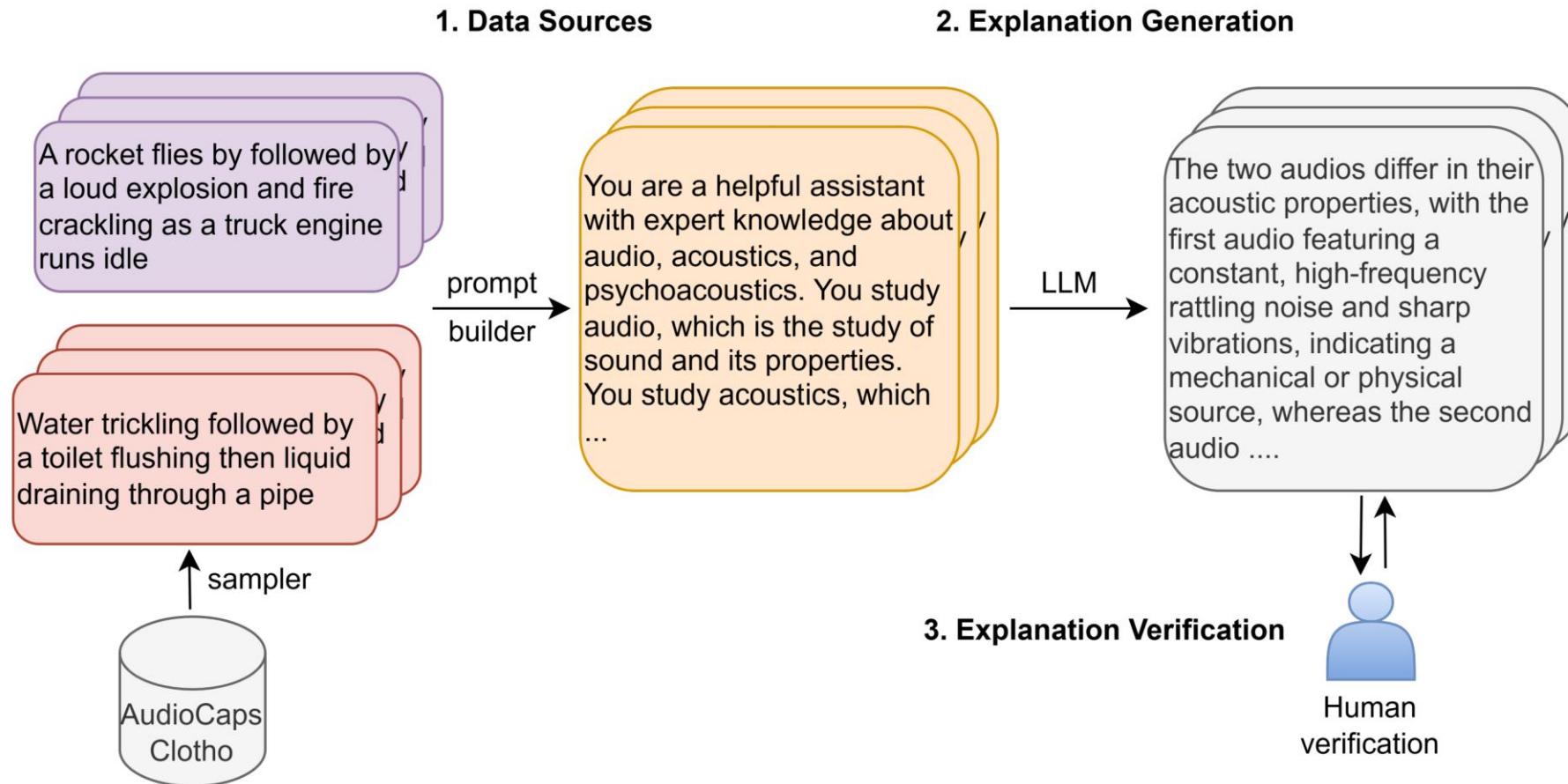
Audio 1  
Audio 2

**Comparative  
reasoning**

Audio difference  
explanation

ADIFF: Explaining audio difference using natural language, <https://arxiv.org/abs/2502.04476>  
ICLR 2025 Spotlight

# Audio Difference Explanation task dataset



# Three tiers of explanations

## Concise

Audio 1 features a dominant speech signal with occasional belches, while Audio 2 combines speech with a persistent, distracting background noise from the motor and wind.

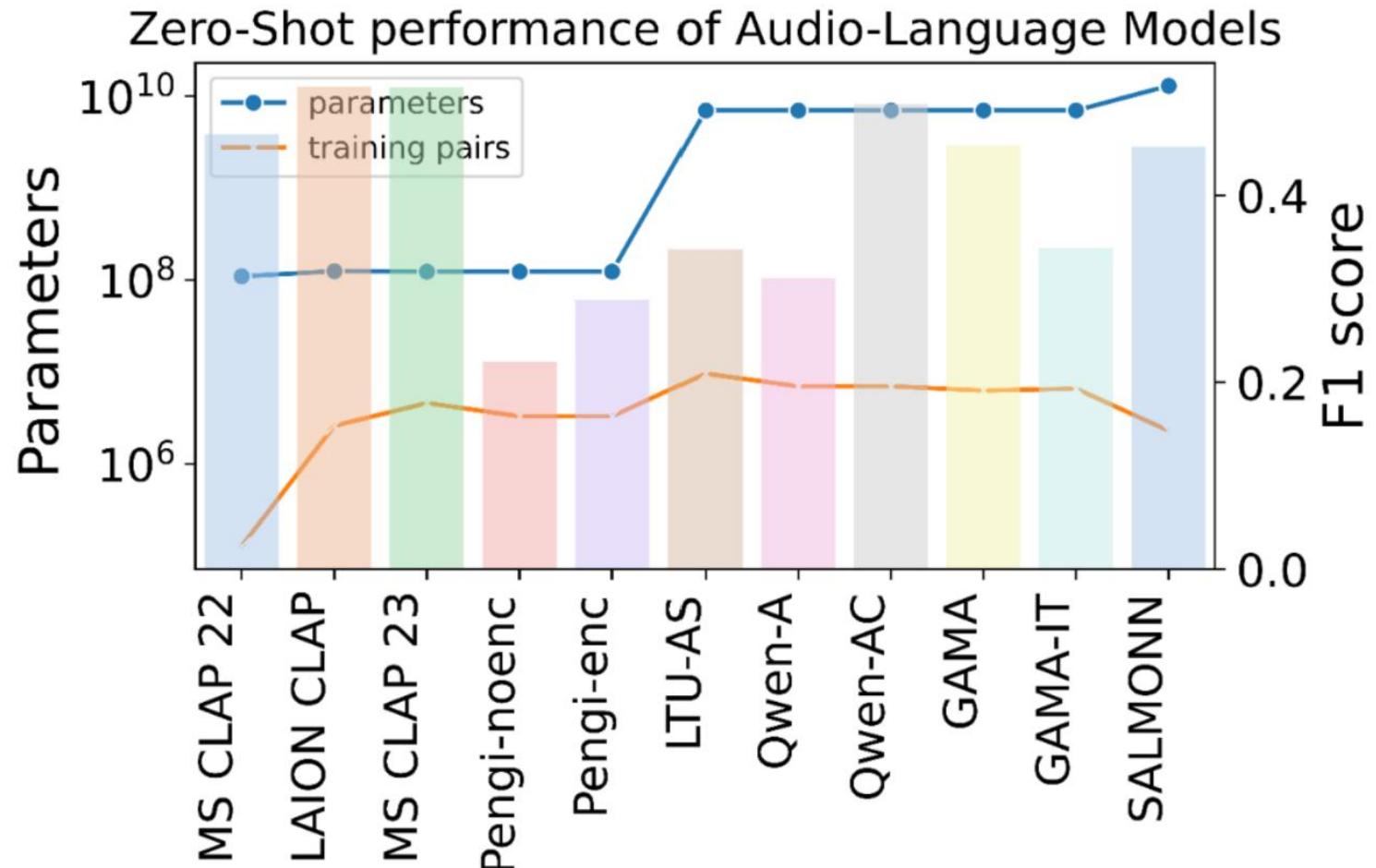
## Brief

The primary difference between the two audios is that the first audio features a clear and intelligible spoken narrative followed by a sudden and jarring interruption of belches, whereas the second audio combines a spoken narrative with a persistent and subtle background noise of a small motor buzzing and wind blowing, which significantly alters the overall sonic texture and listener's perception.

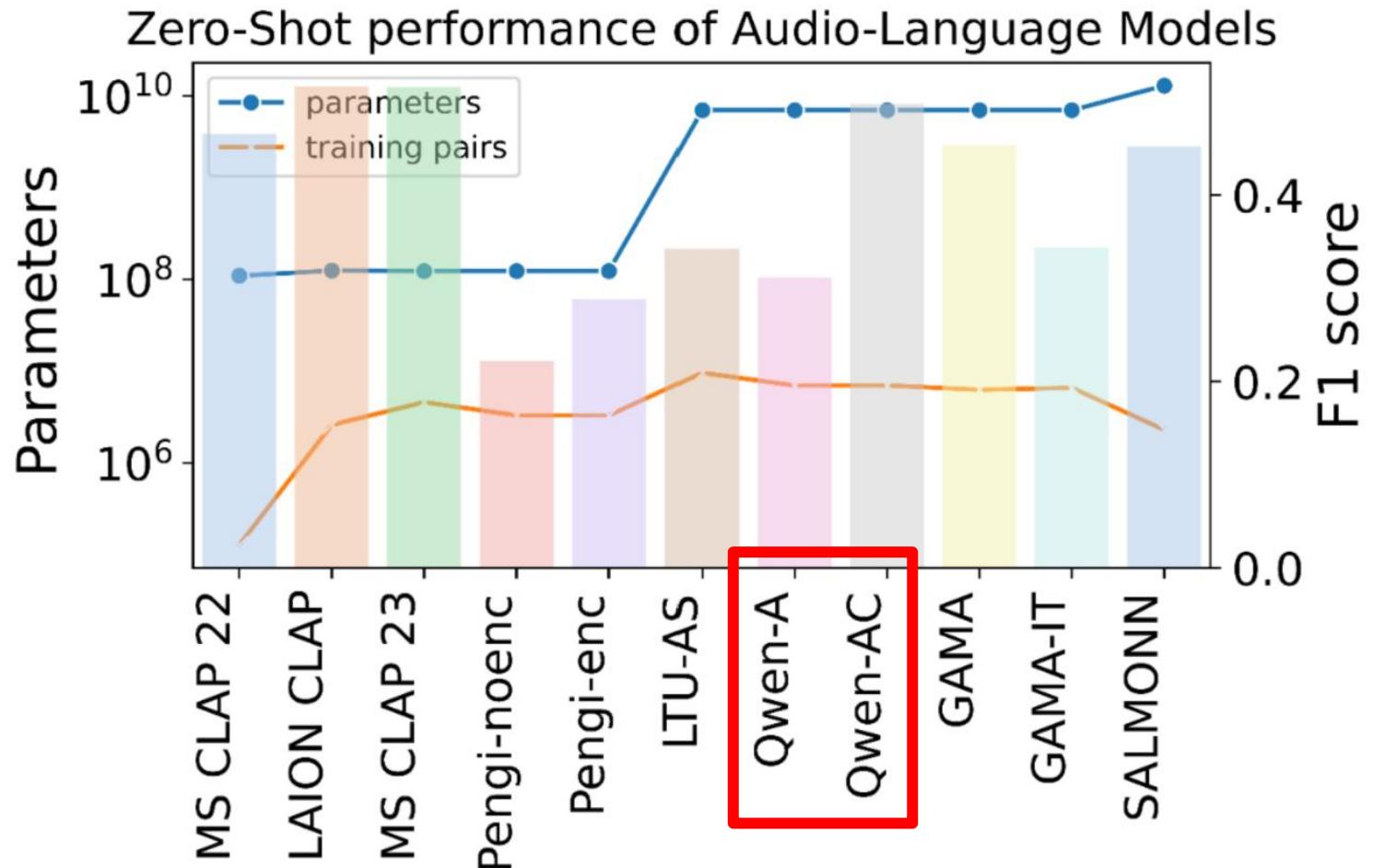
## Detail

As I listen to the two audios, I notice that the first audio, featuring a man talking followed by a series of belches, has a distinct sonic profile. The man's voice has a clear and crisp quality, with a moderate loudness and a frequency range that spans from low to high pitches. The belches that follow have a sharp, sudden onset and a short duration, with a characteristic 'pfft' sound that is high in frequency and low in loudness. In contrast, the second audio features a man speaking with a small motor buzzing and wind blowing, which creates a more complex acoustic environment. The motor's hum has a steady, low-frequency tone that is moderate in loudness, while the wind sounds have a softer, more gentle quality with a higher frequency range. The man's voice is still clear, but it's slightly muffled by the ambient noise of the motor and wind, creating a sense of distance or background. Overall, the first audio is more focused on the man's voice and the belches, while the second audio is more about the interplay between the man's voice and the surrounding environmental sounds.

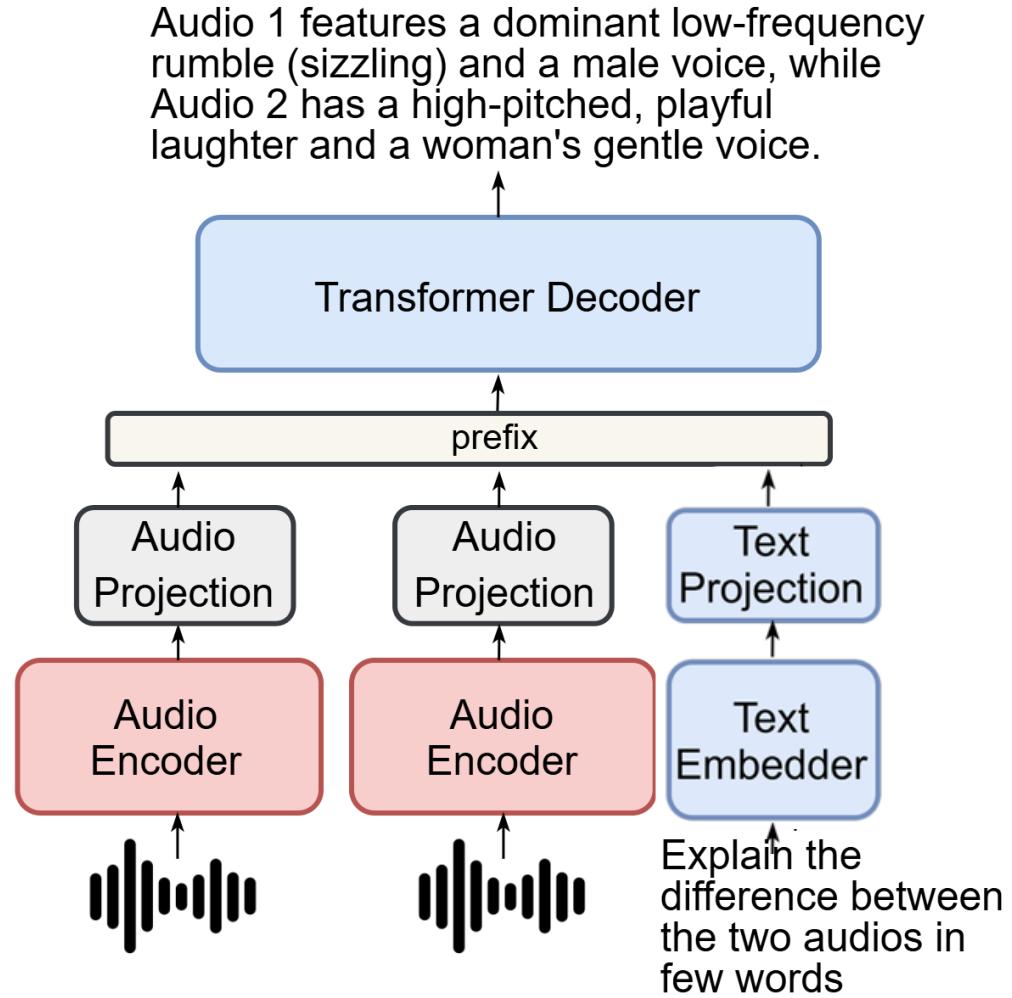
# Unlike entailment benchmarking



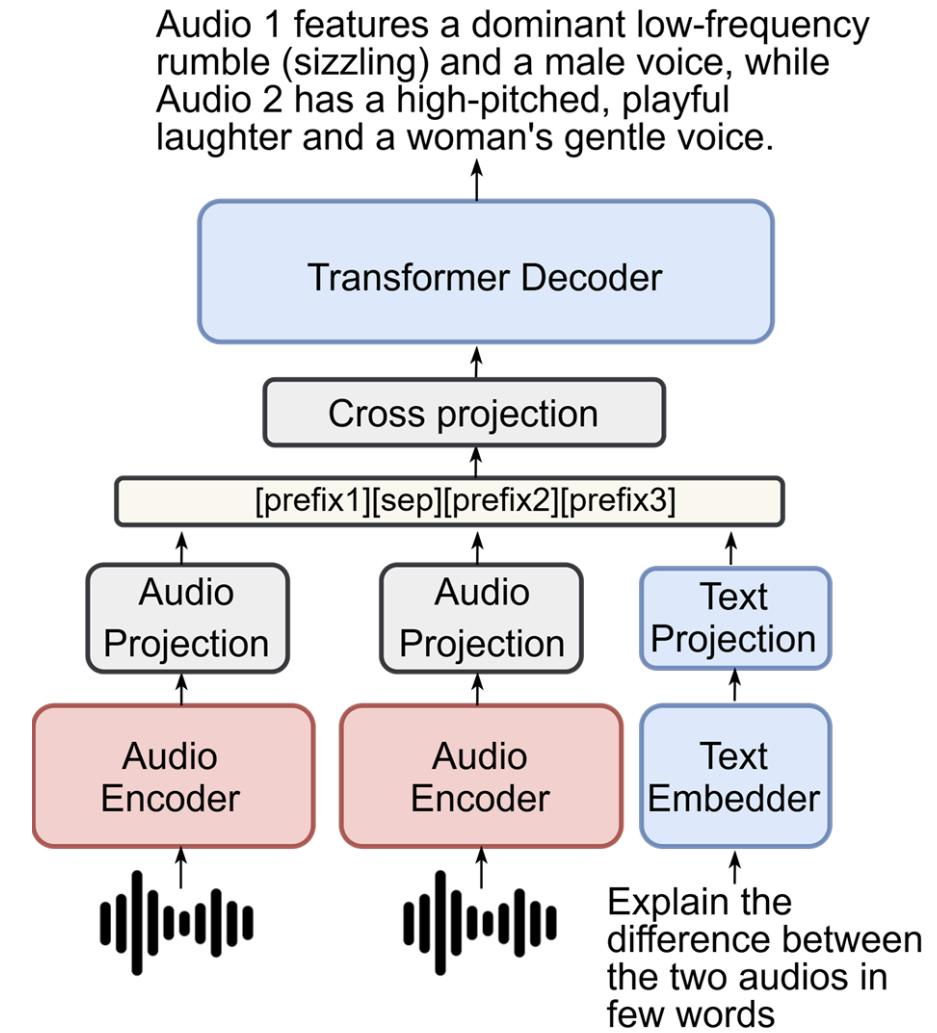
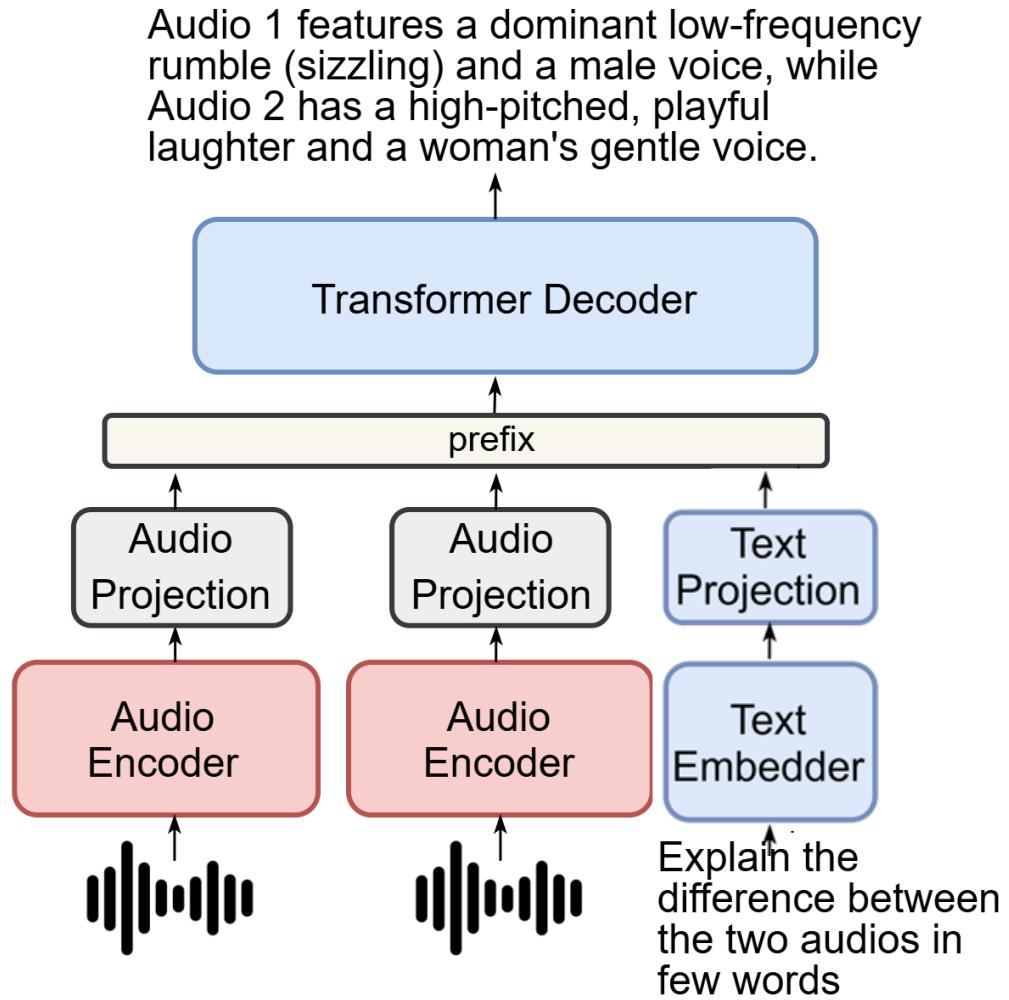
# Only Qwen supports multiple audio input



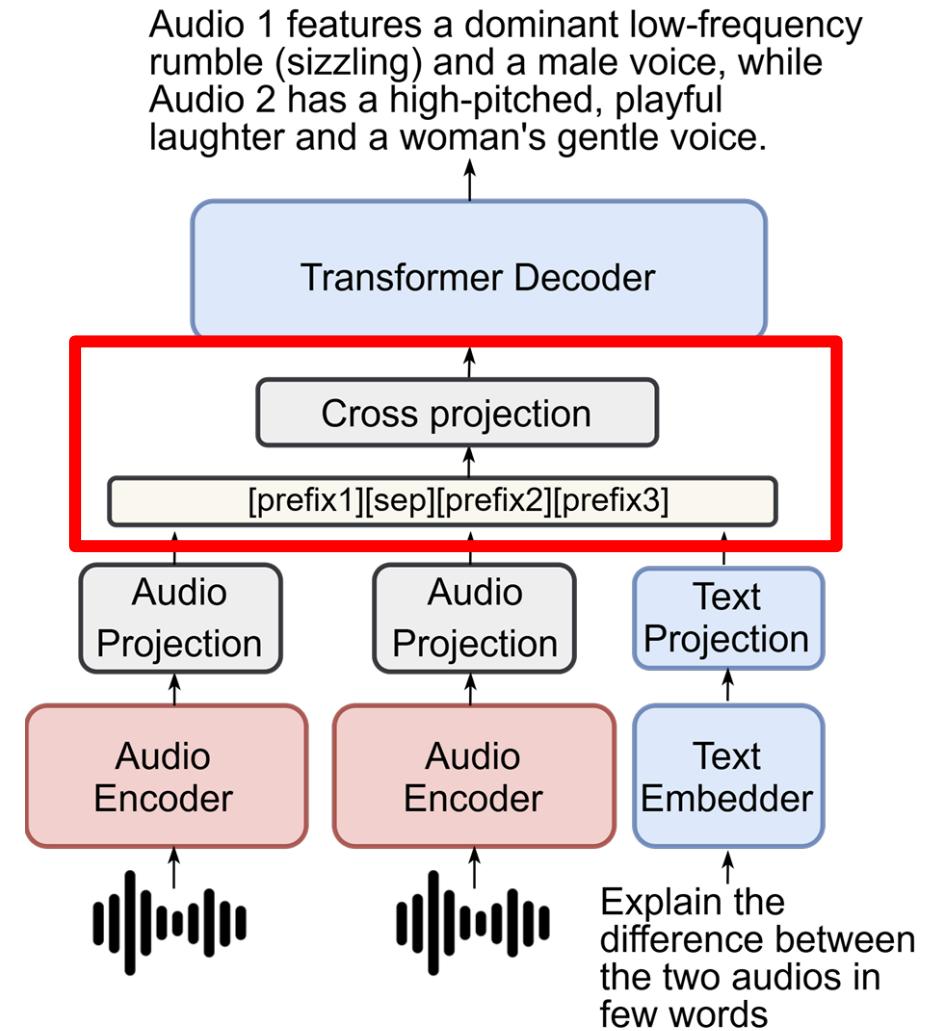
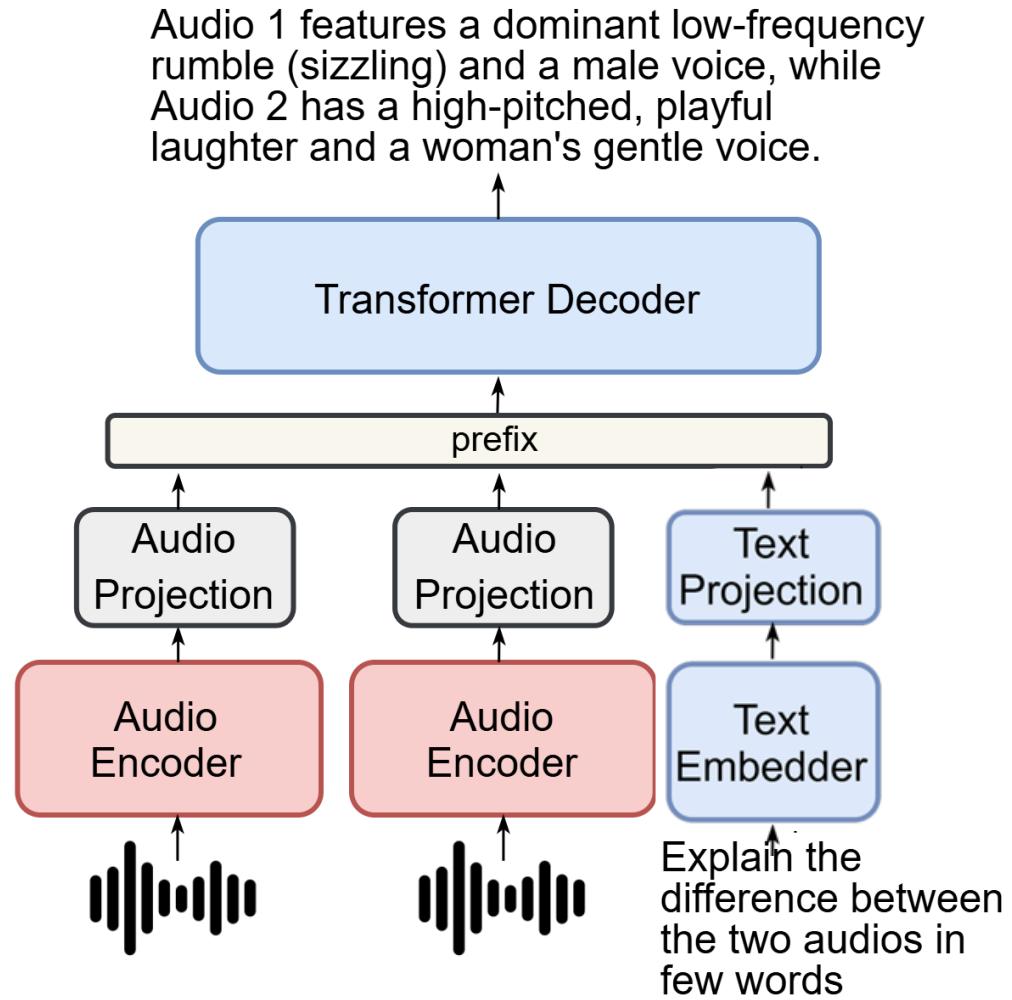
# Introduce naïve baseline



# Proposed ADIFF



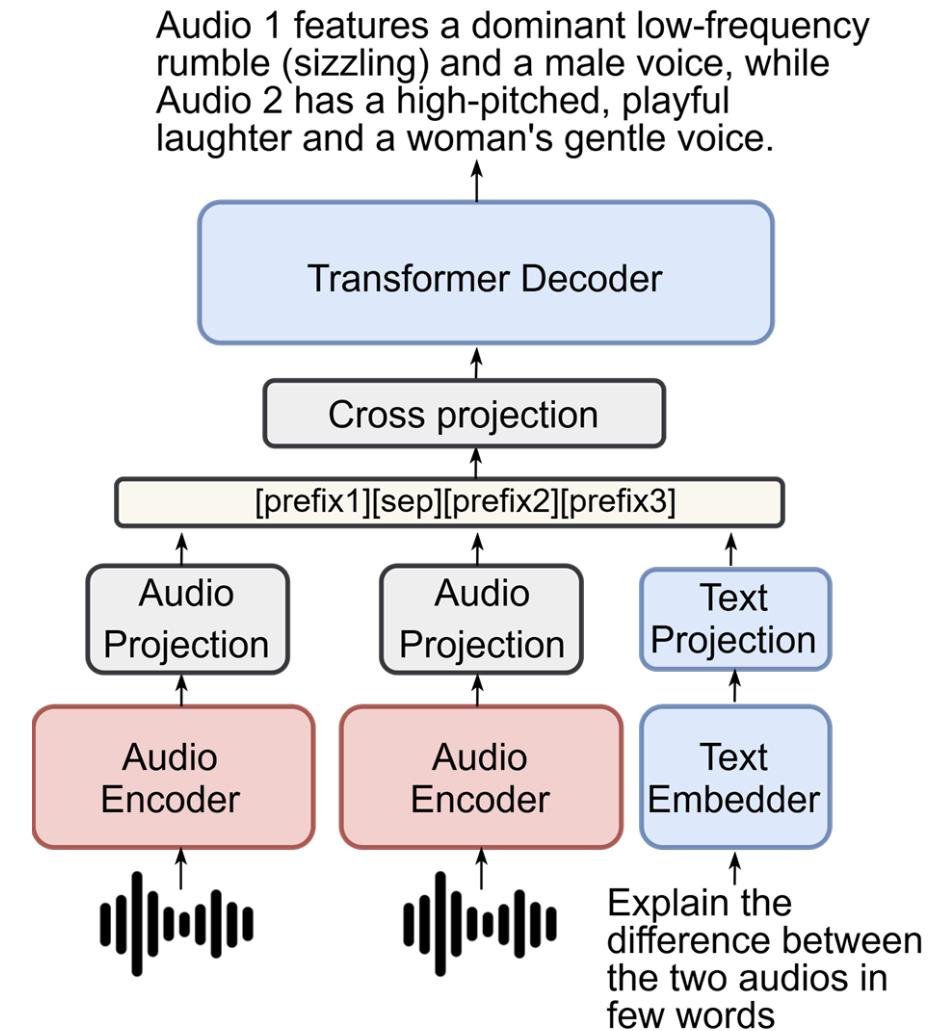
# Proposed ADIFF



# Training ADIFF

- Unimodal pretraining
- Multimodal grounding
- Finetuning

$$\mathcal{L} = - \sum_{i=1}^N \sum_{j=1}^l \log p_\gamma(o_j^i | p_1^i, \dots, p_k^i, o_1^i, \dots, o_{j-1}^i)$$

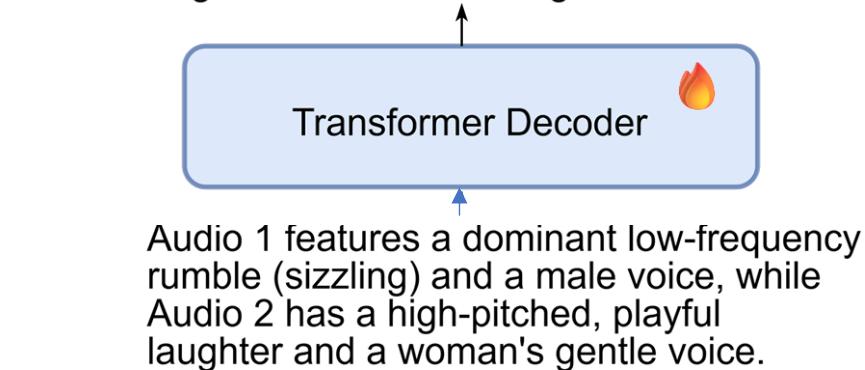


# Training ADIFF

- ✓ Unimodal pretraining
  - Multimodal grounding
  - Finetuning

$$\mathcal{L} = - \sum_{i=1}^N \sum_{j=1}^l \log p_\gamma(o_j^i | p_1^i, \dots, p_k^i, o_1^i, \dots, o_{j-1}^i)$$

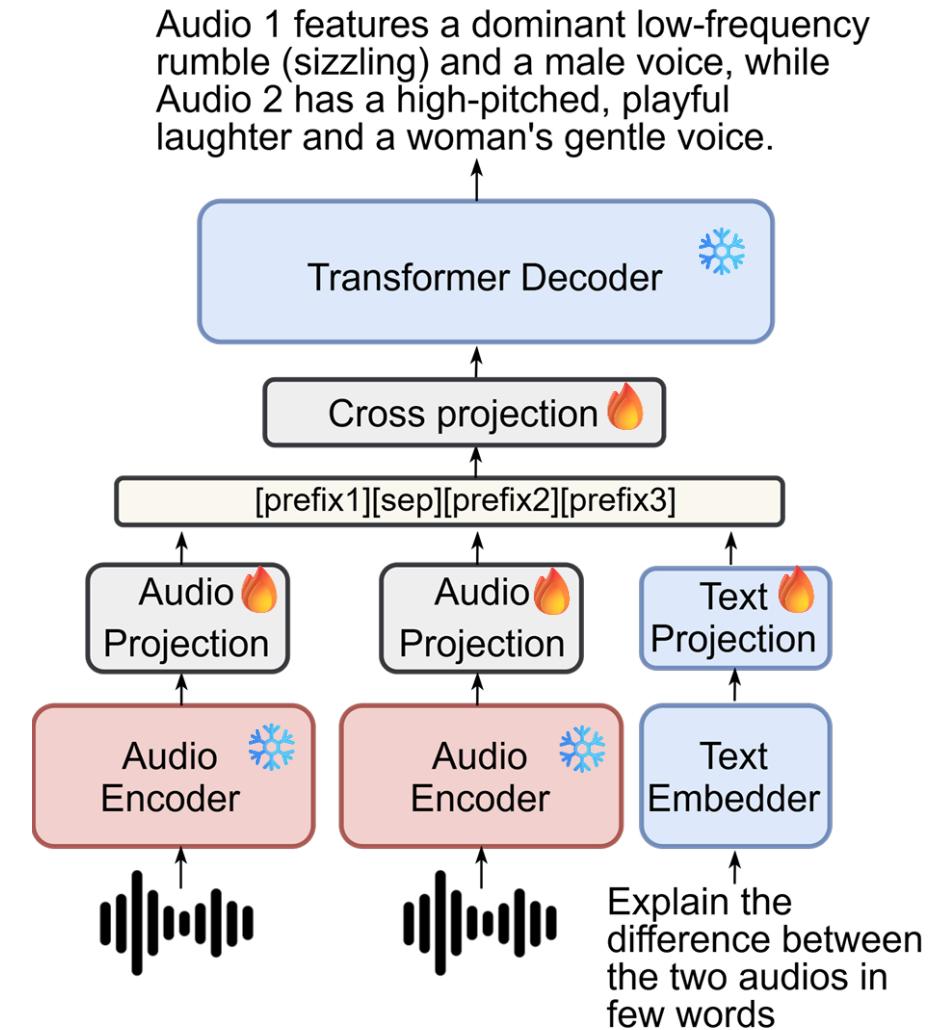
Audio 1 features a dominant low-frequency rumble (sizzling) and a male voice, while Audio 2 has a high-pitched, playful laughter and a woman's gentle voice.



# Training ADIFF

- ✓ Unimodal pretraining
- ✓ Multimodal grounding
  - Finetuning

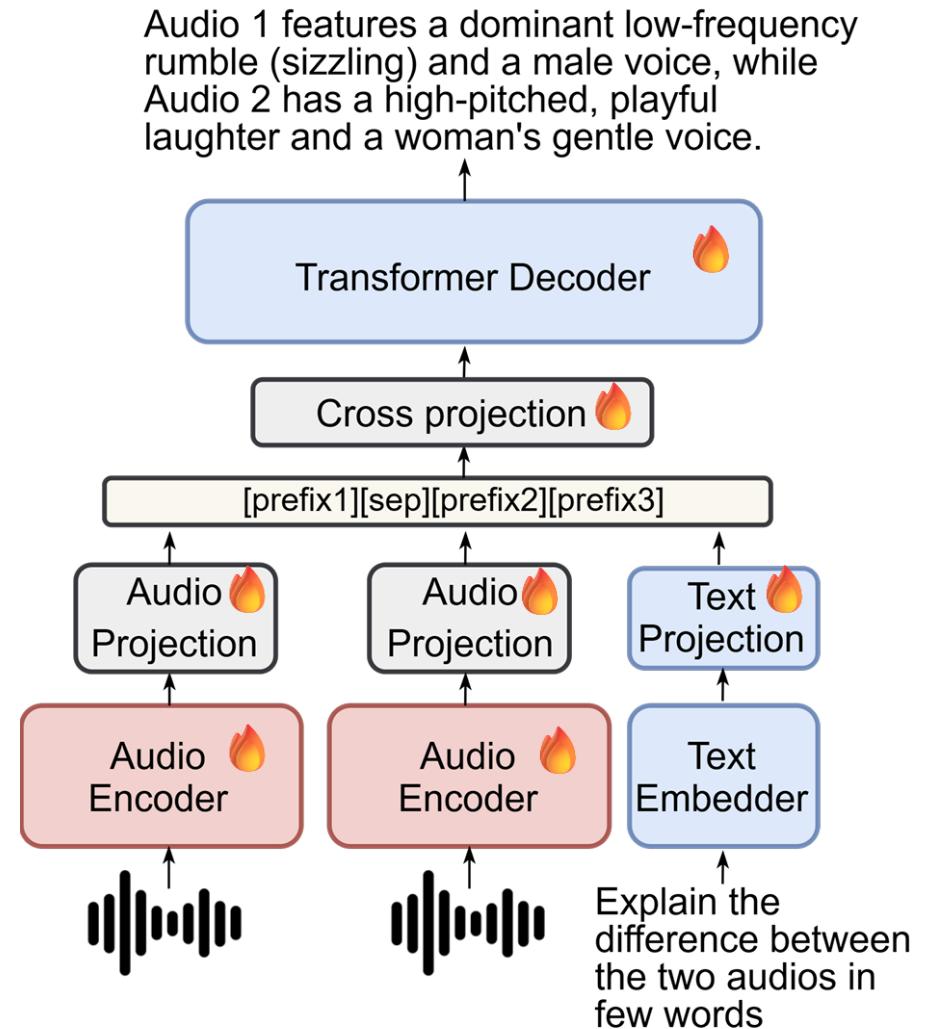
$$\mathcal{L} = - \sum_{i=1}^N \sum_{j=1}^l \log p_\gamma(o_j^i | p_1^i, \dots, p_k^i, o_1^i, \dots, o_{j-1}^i)$$



# Training ADIFF

- ✓ Unimodal pretraining
- ✓ Multimodal grounding
- ✓ Finetuning

$$\mathcal{L} = - \sum_{i=1}^N \sum_{j=1}^l \log p_\gamma(o_j^i | p_1^i, \dots, p_k^i, o_1^i, \dots, o_{j-1}^i)$$



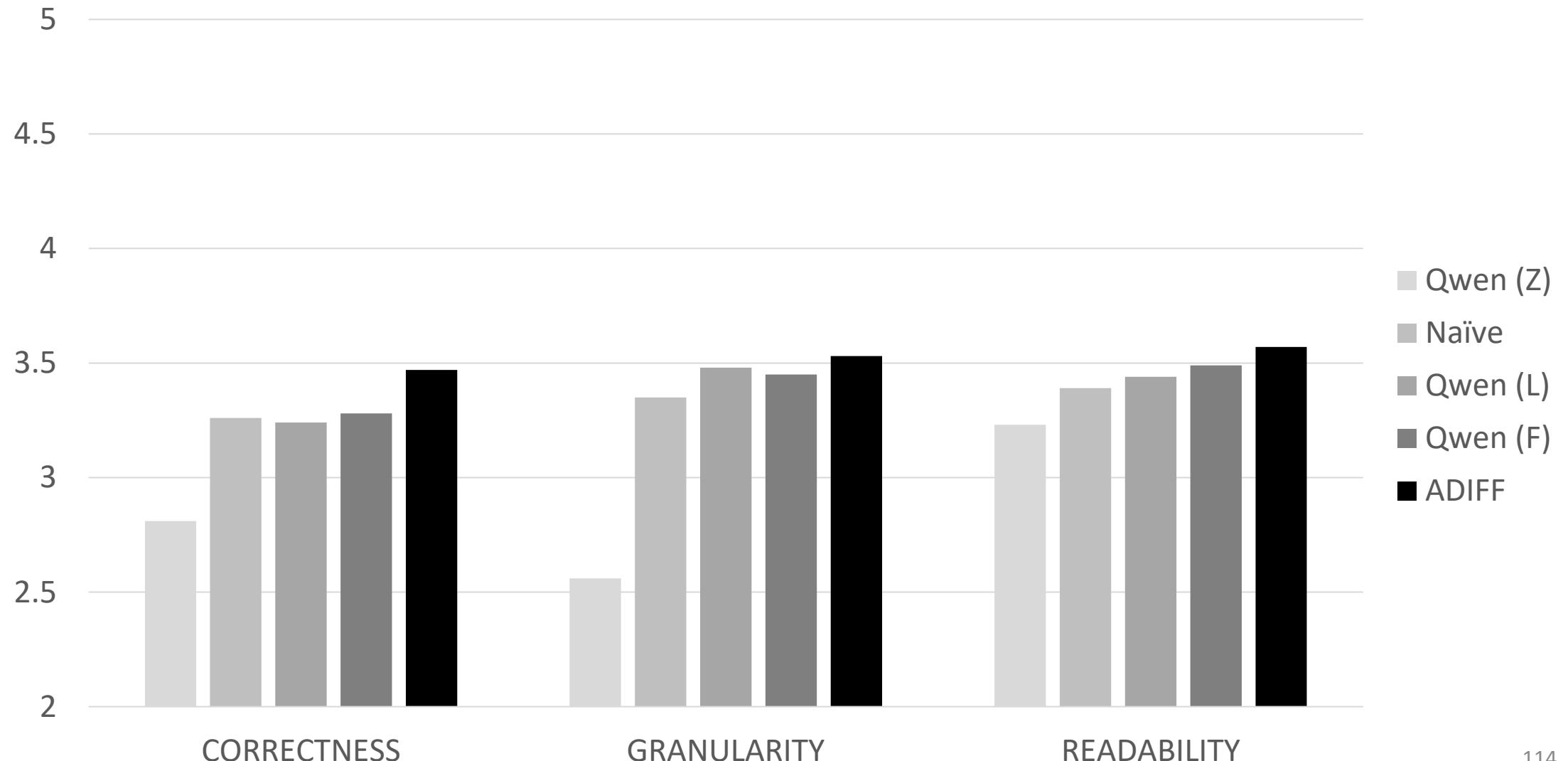
# Results – objective evaluation

Task	Models	Tier 1			Tier 2			Tier 3		
		BLEU <sub>4</sub>	METEOR	SPIDER	BLEU <sub>4</sub>	METEOR	SPIDER	BLEU <sub>4</sub>	METEOR	SPIDER
ACD	Baseline	0.118	0.210	0.220	0.163	0.193	0.225	0.153	0.188	0.123
	QwenAC (L)	0.132	0.214	0.235	0.166	<b>0.204</b>	0.212	0.165	0.202	0.173
	QwenAC (F)	0.110	0.183	0.258	0.163	0.199	0.241	0.151	0.194	0.082
	ADIFF	<b>0.135</b>	<b>0.221</b>	<b>0.303</b>	<b>0.180</b>	0.197	<b>0.345</b>	<b>0.171</b>	<b>0.208</b>	<b>0.183</b>
CLD	Baseline	0.128	0.237	0.212	0.233	0.234	0.641	0.157	0.199	0.166
	QwenAC (L)	0.140	0.285	0.230	0.232	0.236	0.756	0.155	0.200	0.182
	QwenAC (F)	0.126	0.232	0.204	<b>0.273</b>	<b>0.254</b>	<b>0.958</b>	0.130	0.207	0.172
	ADIFF	<b>0.203</b>	<b>0.302</b>	<b>0.652</b>	0.213	0.235	0.692	<b>0.191</b>	<b>0.220</b>	<b>0.417</b>

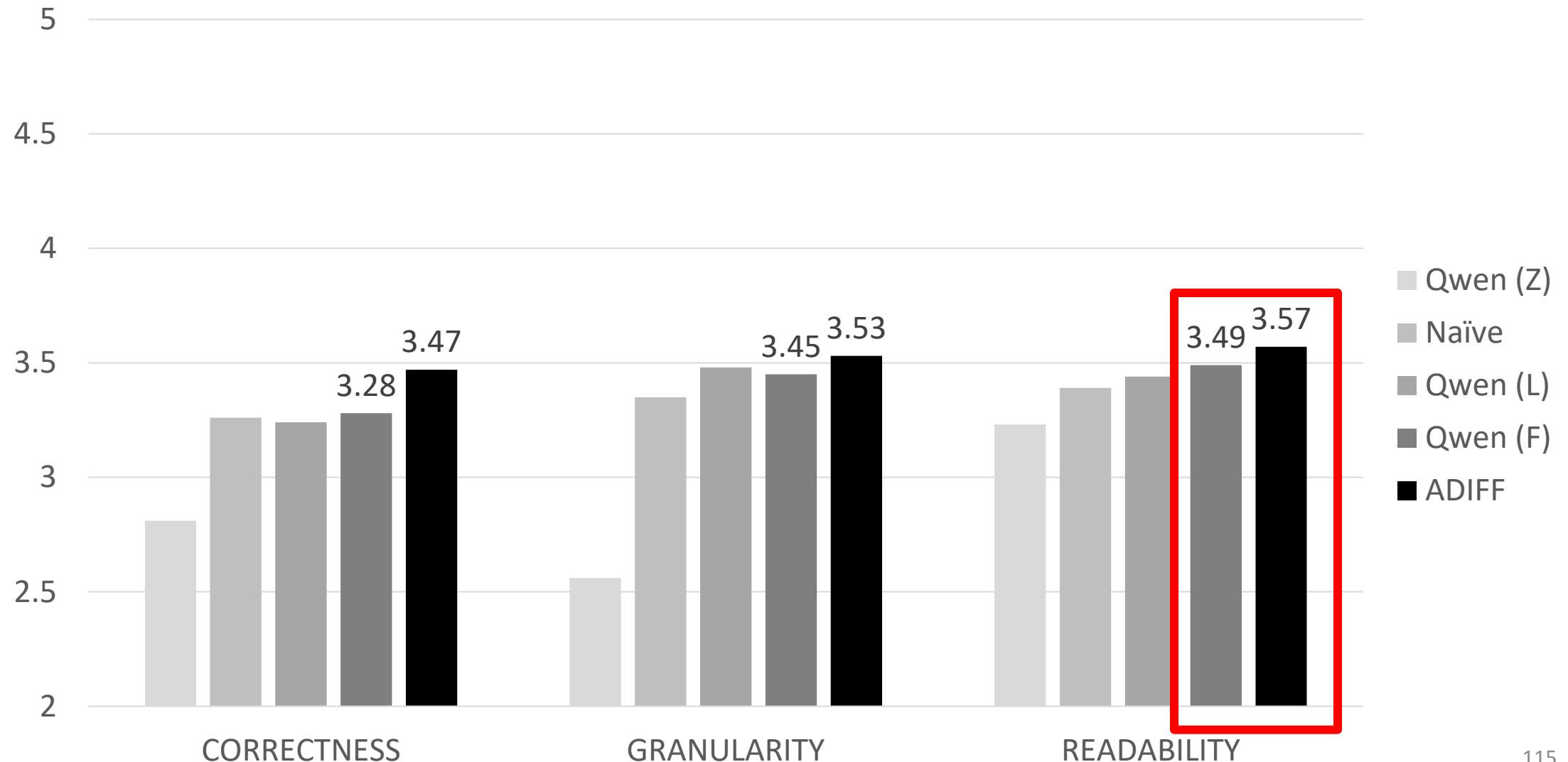
# Results – objective evaluation

Task	Models	Tier 1			Tier 2			Tier 3		
		BLEU <sub>4</sub>	METEOR	SPIDER	BLEU <sub>4</sub>	METEOR	SPIDER	BLEU <sub>4</sub>	METEOR	SPIDER
ACD	Baseline	0.118	0.210	0.220	0.163	0.193	0.225	0.153	0.188	0.123
	QwenAC (L)	0.132	0.214	0.235	0.166	<b>0.204</b>	0.212	0.165	0.202	0.173
	QwenAC (F)	0.110	0.183	0.258	0.163	0.199	0.241	0.151	0.194	0.082
	ADIFF	<b>0.135</b>	<b>0.221</b>	<b>0.303</b>	<b>0.180</b>	0.197	<b>0.345</b>	<b>0.171</b>	<b>0.208</b>	<b>0.183</b>
CLD	Baseline	0.128	0.237	0.212	0.233	0.234	0.641	0.157	0.199	0.166
	QwenAC (L)	0.140	0.285	0.230	0.232	0.236	0.756	0.155	0.200	0.182
	QwenAC (F)	0.126	0.232	0.204	<b>0.273</b>	<b>0.254</b>	<b>0.958</b>	0.130	0.207	0.172
	ADIFF	<b>0.203</b>	<b>0.302</b>	<b>0.652</b>	0.213	0.235	0.692	<b>0.191</b>	<b>0.220</b>	<b>0.417</b>

# Results – subjective evaluation



# Results – subjective evaluation



# Ablations

		Tier 1			Tier 2			Tier 3		
Task	Exp.	BLEU <sub>4</sub>	METEOR	SPIDER	BLEU <sub>4</sub>	METEOR	SPIDER	BLEU <sub>4</sub>	METEOR	SPIDER
ACD	A	0.082	0.184	0.128	0.135	0.176	0.154	0.149	0.175	0.107
	B	0.118	0.210	0.220	0.163	0.193	0.225	0.199	0.188	0.123
	C	0.131	0.214	0.287	0.155	0.197	0.243	0.149	0.203	0.154
	C'	0.139	0.220	0.289	0.186	0.201	0.337	0.175	0.211	0.172
	D	0.129	0.208	0.300	0.162	0.195	0.296	0.164	0.199	0.150
	E	0.135	0.221	0.303	0.180	0.197	0.345	0.171	0.208	0.183
CLD	A	0.138	0.223	0.127	0.199	0.236	0.589	0.123	0.166	0.127
	B	0.128	0.237	0.212	0.233	0.234	0.641	0.157	0.199	0.166
	C	0.156	0.257	0.195	0.280	0.267	0.904	0.127	0.187	0.196
	C'	0.210	0.305	0.639	0.289	0.271	0.906	0.187	0.216	0.405
	D	0.191	0.304	0.511	0.236	0.242	0.689	0.136	0.188	0.124
	E	0.203	0.302	0.652	0.213	0.235	0.692	0.191	0.220	0.417

Table 2: Ablation results. Experiment A is baseline architecture with random audio encoder weights. Experiment B is pretrained audio encoder weights. Experiment C is ADIFF which modifies baseline architecture with separator token and cross-projection. Experiment C' applies the LM finetuning to the ADIFF model from Experiment C. Experiment D is the ADIFF model with position captioning. Experiment E finetunes the language model of ADIFF along with position captioning.

# Ablation findings

1. Cross-projection aids in utilizing text prefixes to store difference attributes

---

## Difference explanation from ADIFF

---

As I listen to the two audios, I notice that Audio (1) features a sudden, sharp, and high-frequency explosion followed by a series of crackling sounds, likely generated by a small object being hit. The explosion's frequency content is likely dominated by mid-to-high frequencies, with a prominent presence of low-frequency energy. In contrast, Audio (2) presents a continuous, low-frequency hum of a helicopter flying overhead, with a consistent pitch and amplitude. The helicopter's sound is likely generated by a series of high-frequency transients, with a prominent presence of mid-to-high frequencies. The overall dynamic range of Audio (2) is much greater, with sudden changes in loudness and frequency content, whereas Audio (1) has a more gradual onset and decay.

---

### Audio 1 prefix

---

explosion, crack, hit

---

---

### Audio 2 prefix

---

helicop, loud

---

---

### Text prefix

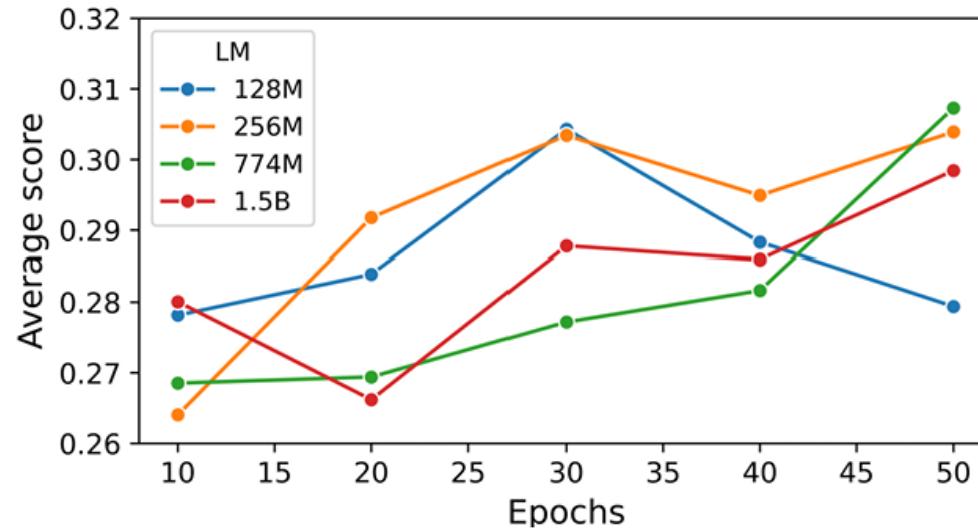
---

mid-high, frequency, pitch, dynamic, range

---

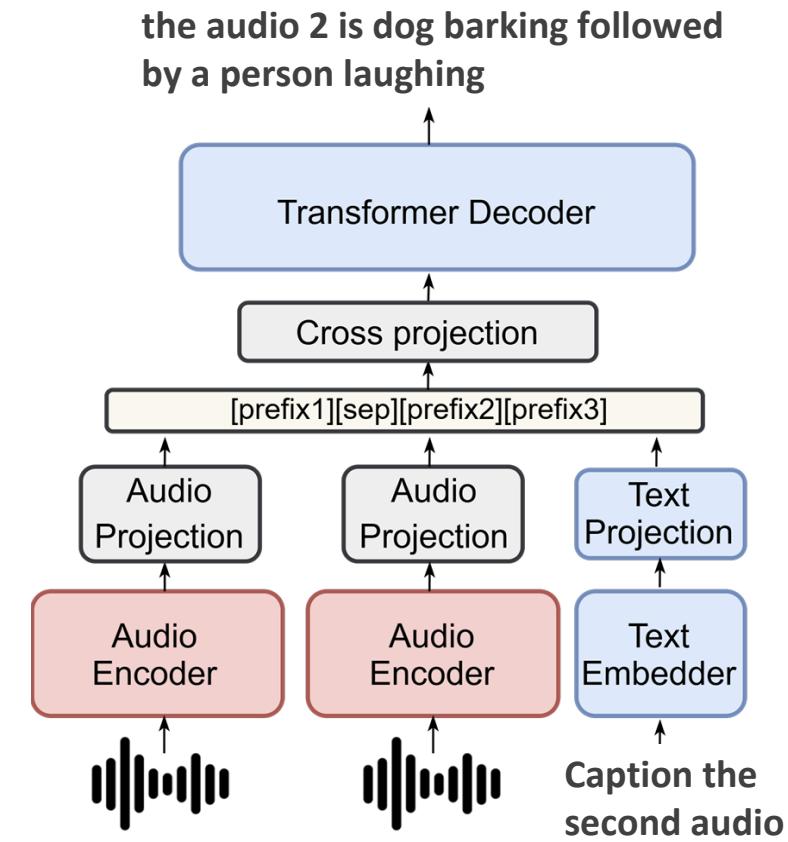
# Ablation findings

1. Cross-projection aids in utilizing text prefixes to store difference attributes
2. Under limited compute and data, smaller LMs are easier to ground in audio with proper training



# Ablation findings

1. Cross-projection aids in utilizing text prefixes to store difference attributes
2. Under limited compute and data, smaller LMs are easier to ground in audio with proper training
3. Position-based audio captioning improves model performance on similar-sounding acoustic sources



# Talk outline

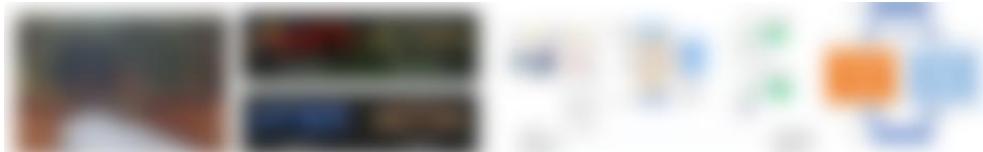
- Effective learning and pretraining methods  
**CLAP (ICASSP 23, 24), Pengi (NeurIPS 23), Mellow (under review)**



- Methods or benchmarks to analyze new abilities  
**Audio Entail (AAAI 25), Audio Difference (ICLR 25), PAM (INTERSPEECH 24)**



- Efficient adaptation methods to enable customization  
**Test-time (ICASSP 24), Few-shot (INTERSPEECH 25), Missing modality (ICASSP 25)**



# Talk outline

- Effective learning and pretraining methods  
CLAP (ICASSP 23, 24), Pengi (NeurIPS 23), Mellow (under review)



- Methods or benchmarks to analyze new abilities  
Audio Entail (AAAI 25), Audio Difference (ICLR 25), PAM (INTERSPEECH 24)



- **Efficient adaptation methods to enable customization**

Test-time (ICASSP 24), Few-shot (INTERSPEECH 25), Missing modality (ICASSP 25)

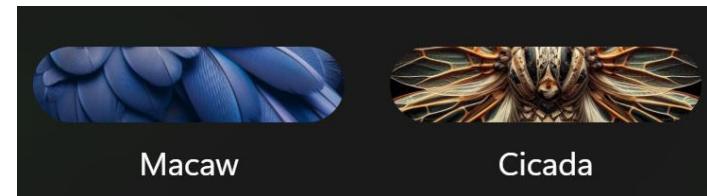
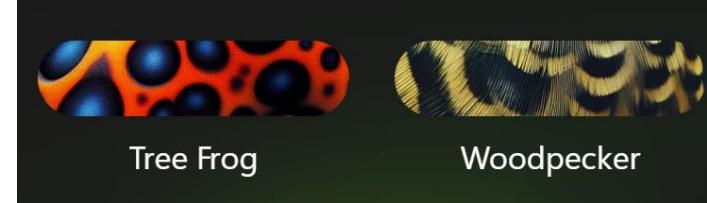


# Adaptation methods

Audio Foundation Models are trained on millions of audio-text pairs. However, their performance across domains is not uniform.

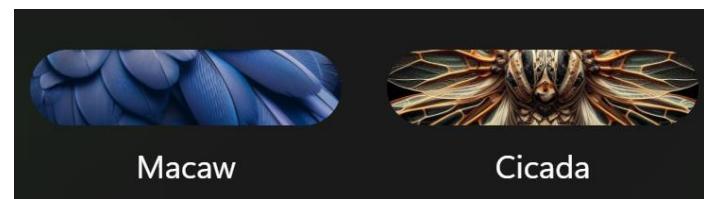
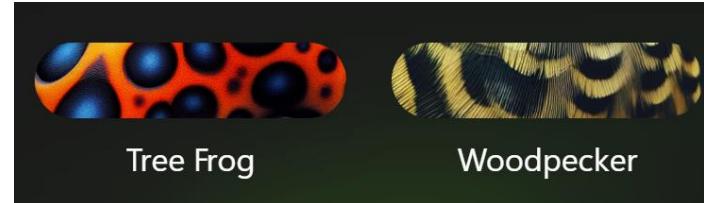
# Adaptation methods

Audio Foundation Models are trained on millions of audio-text pairs. However, their performance across domains is not uniform.



# Adaptation methods

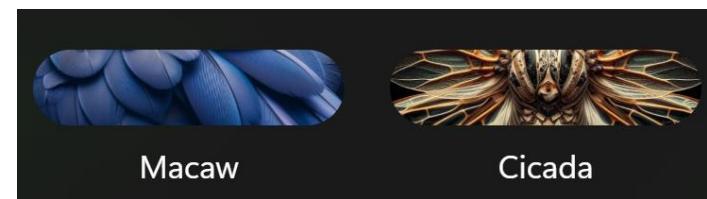
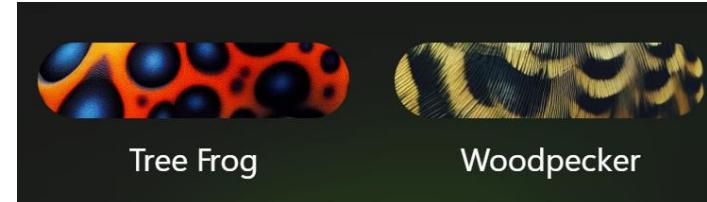
Audio Foundation Models are trained on millions of audio-text pairs. However, their performance across domains is not uniform.



**Performance on OOD scenarios is near random**

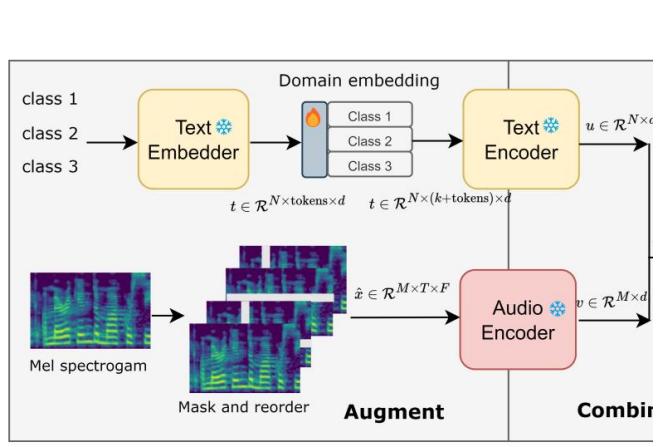
# Adaptation methods

Audio Foundation Models are trained on millions of audio-text pairs. However, their performance across domains is not uniform.

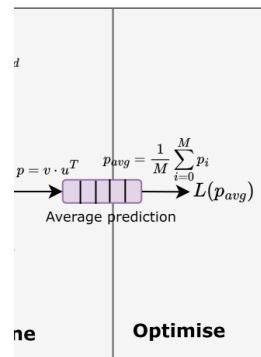


**Adapting Audio Foundation Models is necessary for OOD scenarios**

# Adaptation methods overview



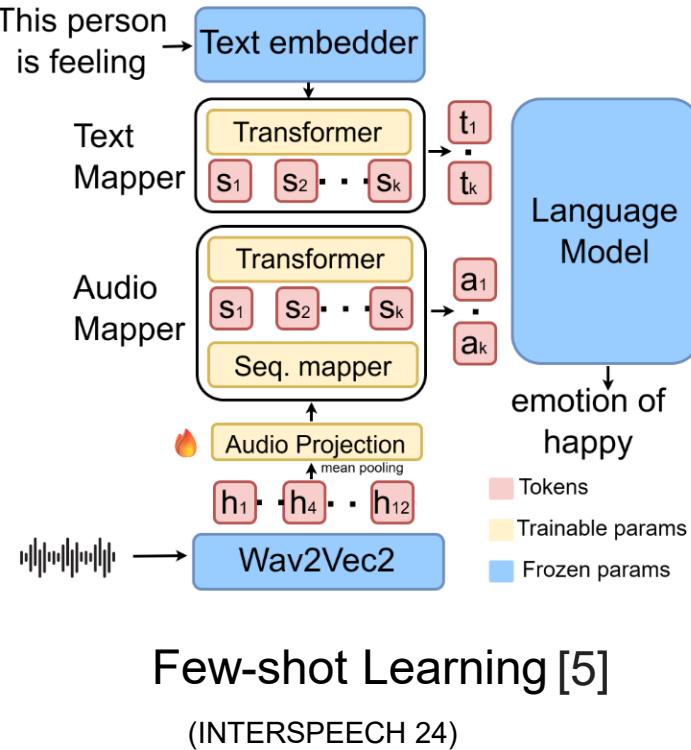
**Test-time Adaptation [4]**  
(INTERSPEECH 24)



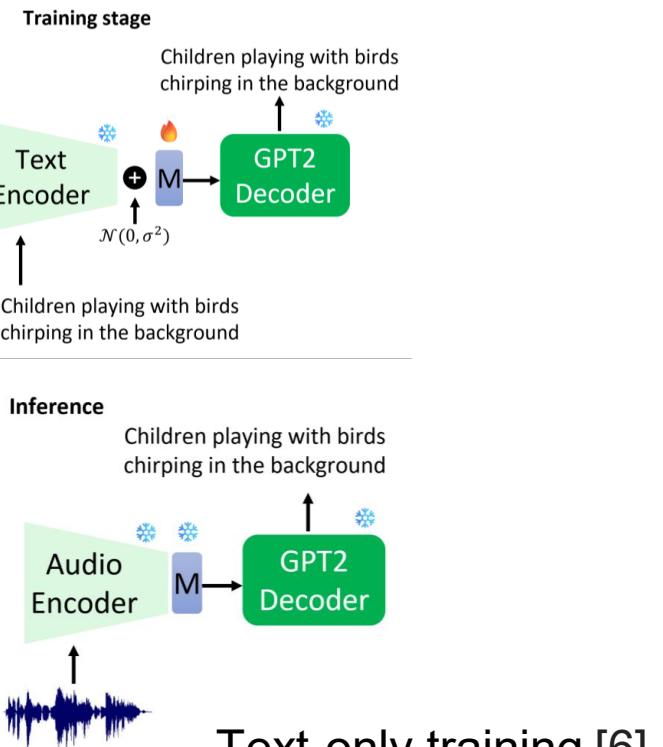
*Test time*

*Least compute*

*Least data*



**Few-shot Learning [5]**  
(INTERSPEECH 24)



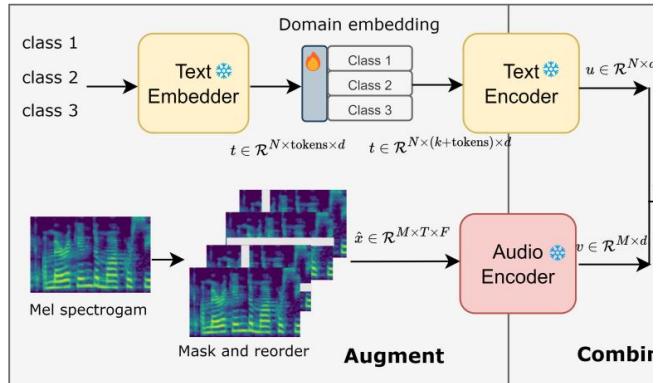
**Text-only training [6]**  
(ICASSP 24)

*Training time*

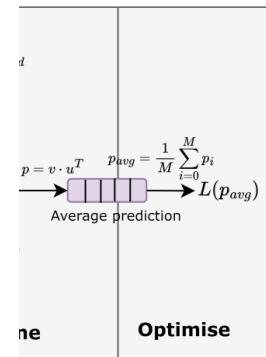
*Most compute*

*Most data*

# Adaptation methods overview



**Test-time  
Adaptation [4]**  
(INTERSPEECH 24)



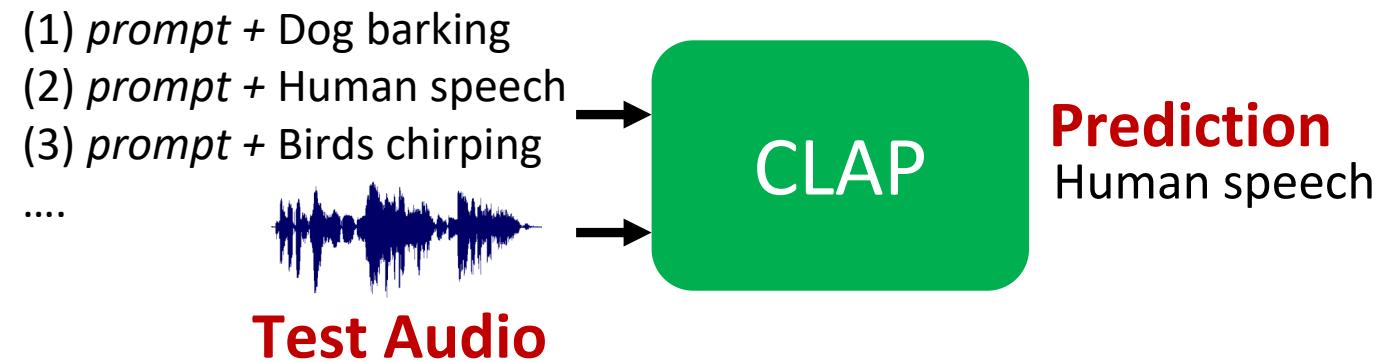
*Test time*

*Least compute*

*Least data*

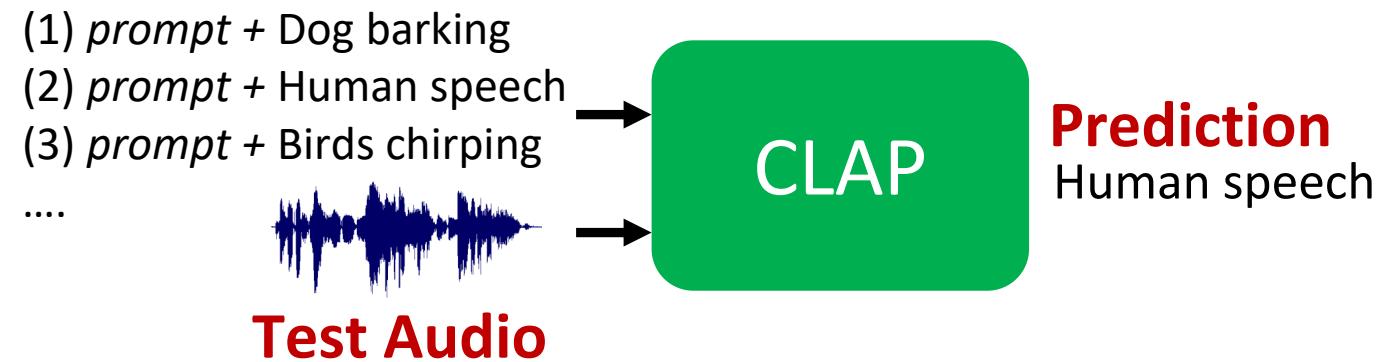
# Contrastive Audio-Language Models

An example zero-shot inference for sound event classification



# Contrastive Audio-Language Models

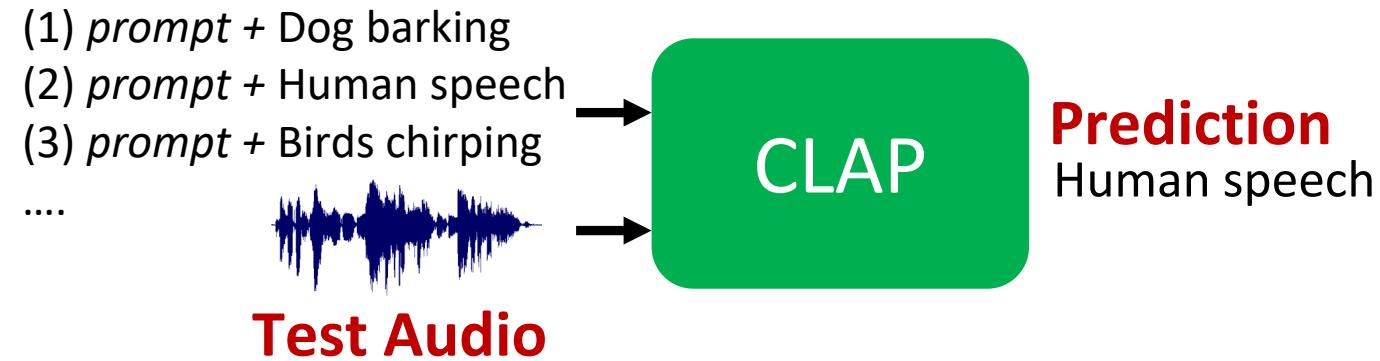
An example zero-shot inference for sound event classification



A prompt of "this is the sound of " is used for Sound Event Classification while "this emotion is " is used for Speech Emotion Recognition

# Contrastive Audio-Language Models

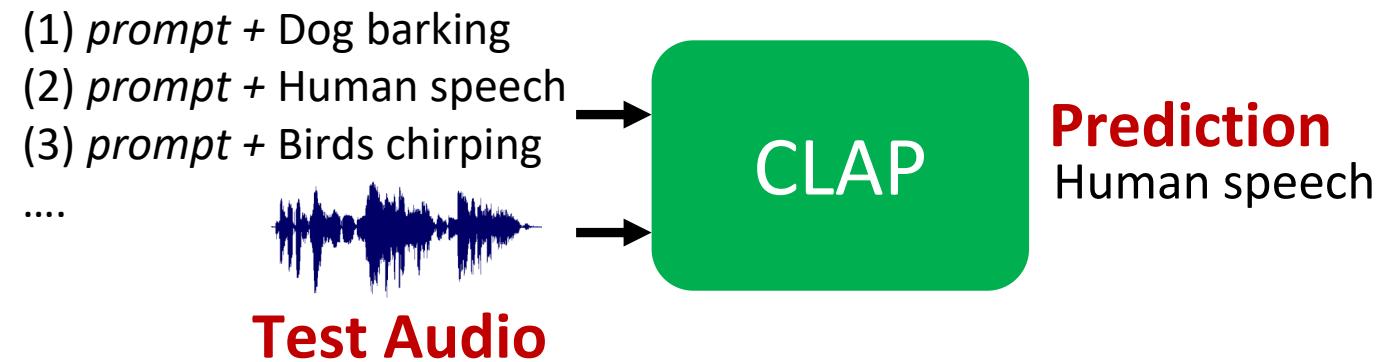
An example zero-shot inference for sound event classification



**The choice of prompt affects zero-shot classification performance (~5%)**

# Contrastive Audio-Language Models

An example zero-shot inference for sound event classification



**Can we automatically learn the prompt or domain vector at test-time instead of hand-crafting them?**

# Test-time domain adaptation for AFM

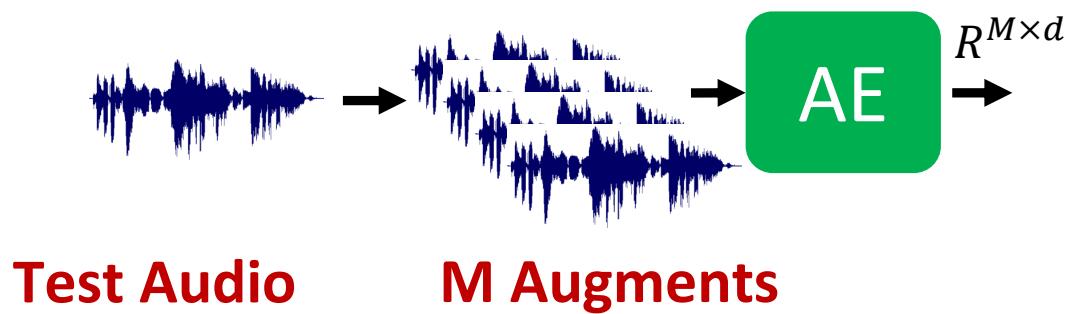
Learn a **domain vector** that **minimizes the self-entropy** of prediction probability obtained from multiple audio augmentations for test audio

**1** Augment

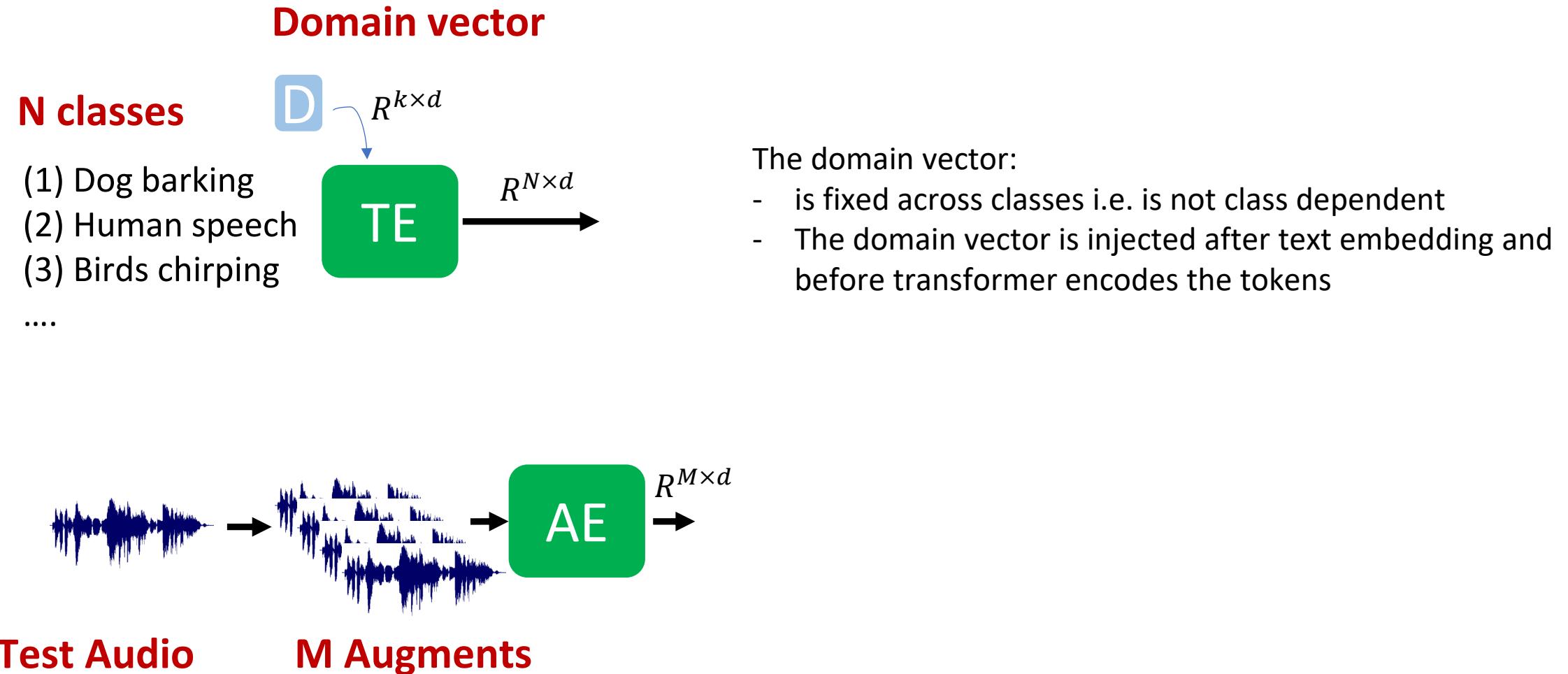
**2** Combine

**3** Optimize

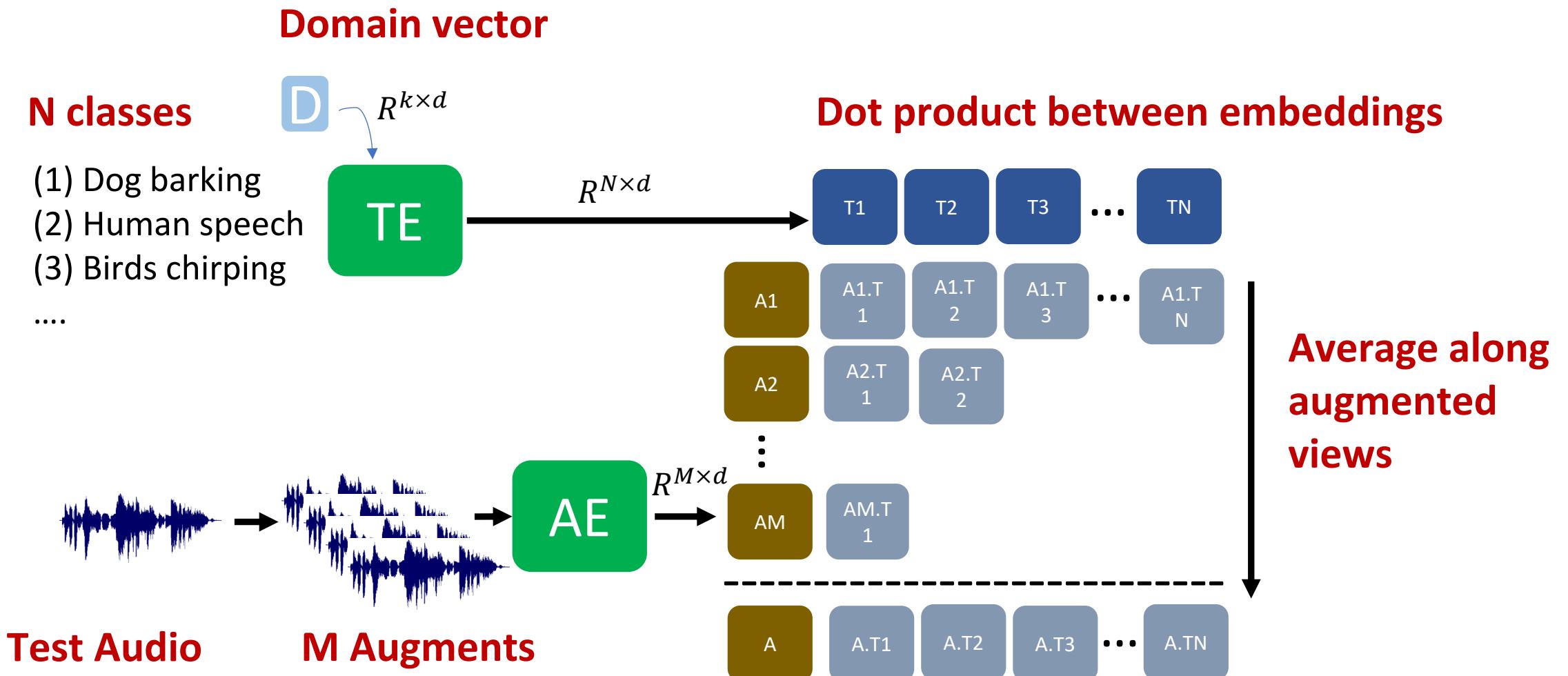
# Test-time domain adaptation: Augment



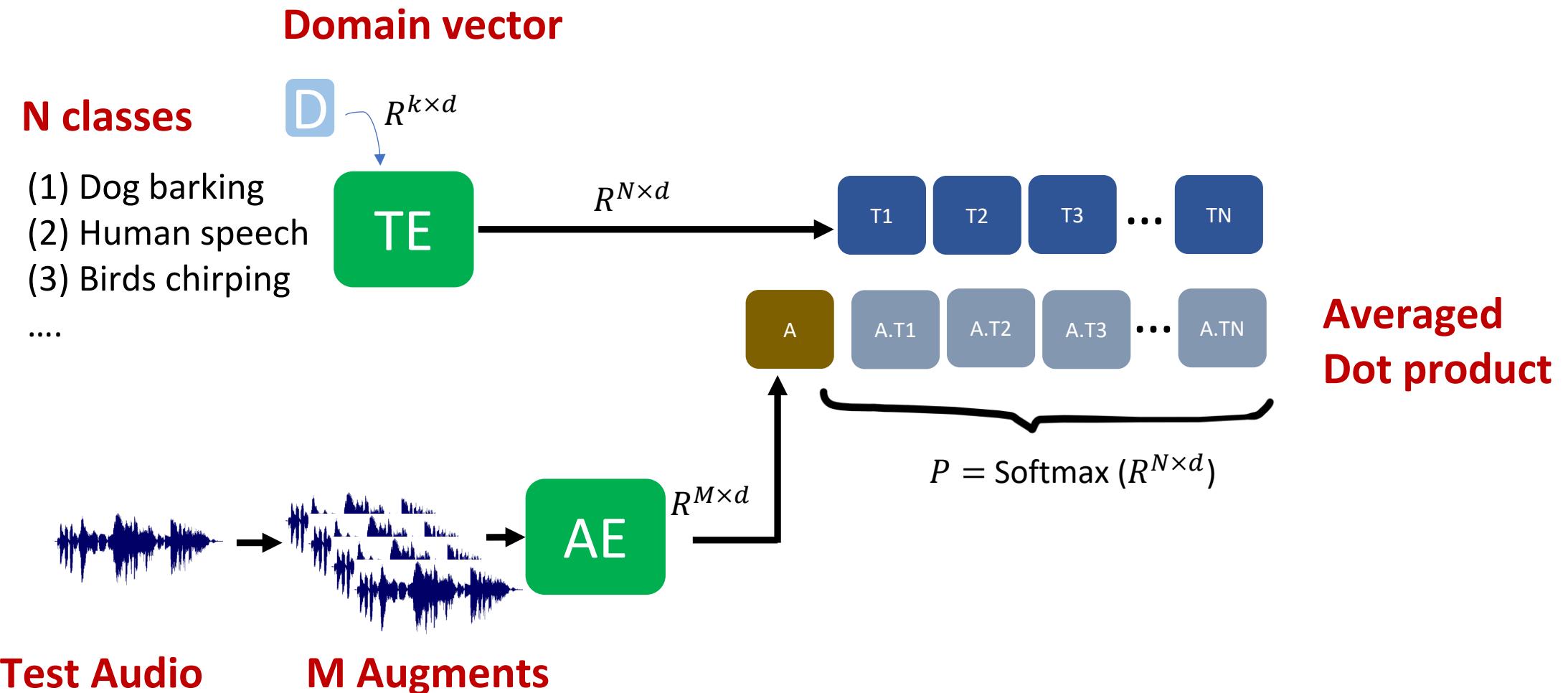
# Test-time domain adaptation: Augment



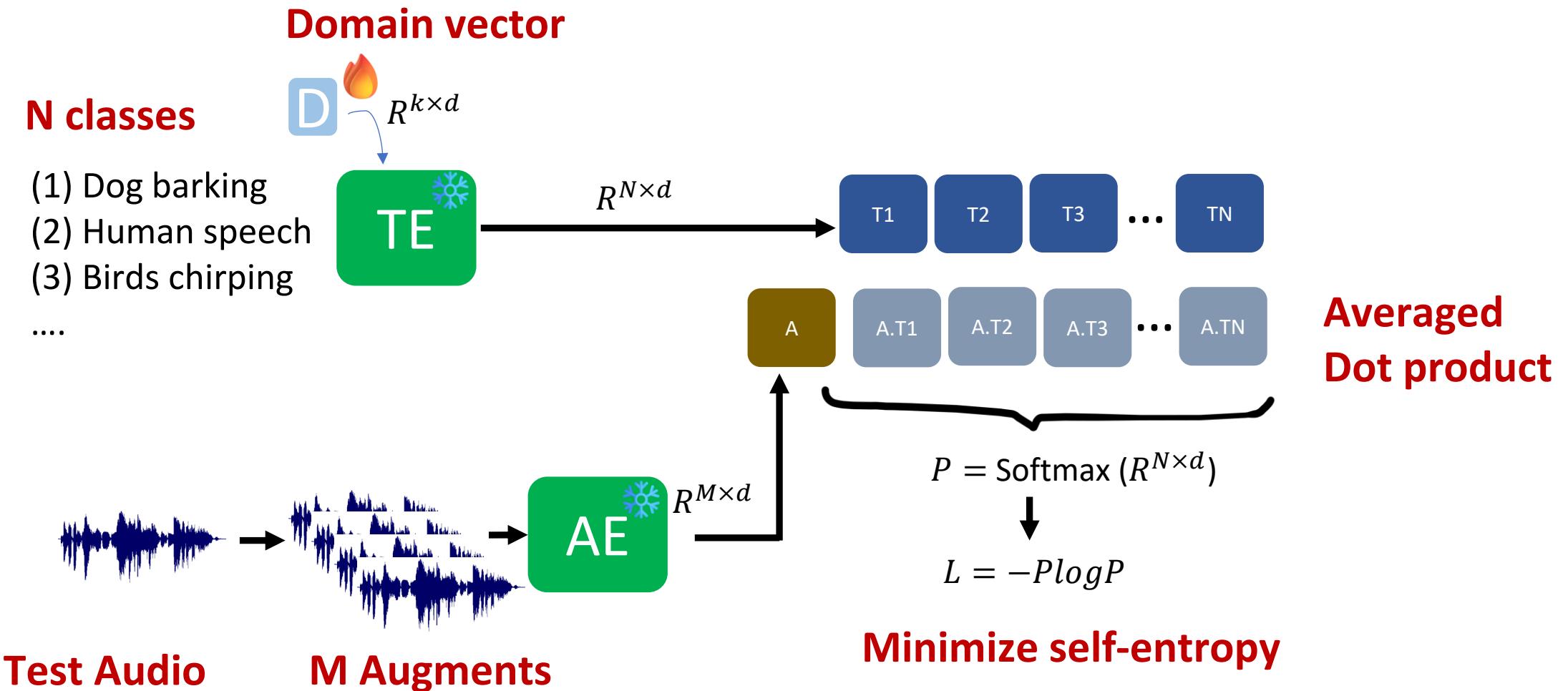
# Test-time domain adaptation: Combine



# Test-time domain adaptation: Combine



# Test-time domain adaptation: Optimize



# Experiments

Various tasks with audio duration ranges from 3 to  $\geq 35$  seconds, classes range from binary to 50 classes, audio files per data range from 120 to 305k files.

Domains: Sound Event Classification, Acoustic Scene Classification, Vocal Sounds, Music, Surveillance, and Speech Emotion Recognition.

Datasets: ESC50, UrbanSound8K, DCASE2017 Task4, TUT 2017, GTZAN Music Speech, GTZAN Genres, Beijing Opera Percussions, CREMA-D, RAVDESS, Vocal Sound

# Results: test-time adapt vs zero-shot baseline

	Average ↑	Sound Event Classification ↑			Vocal Sound Classification ↑	Surveillance Sound Classif.↑	Acoustic Scene Classification↑
Model	Average	ESC50	US8K	DCASE17 Task 4	Vocal Sound	SESA	TUT 2017
Zero-Shot	62.93	93.90	82.30	46.60	79.97	64.95	53.80
One audio	64.94	93.35	<b>85.26</b>	50.96	82.14	73.30	54.19
Five audio	<b>65.92</b>	<b>95.05</b>	85.21	<b>52.30</b>	<b>82.40</b>	<b>74.35</b>	<b>54.38</b>

	Music Classification ↑		Instrument Classification ↑		Speech Emotion Classification↑	
Model	GTZAN Music	GTZAN Speech	Beijing Opera	NS Instr. family	CRE MA-D	RAV DESS
Zero-Shot	99.20	58.40	46.60	68.00	30.00	31.54
One audio	99.21	61.00	47.45	68.28	29.92	<b>34.25</b>
Five audio	<b>100.0</b>	<b>63.20</b>	<b>50.42</b>	<b>69.23</b>	<b>31.27</b>	33.19

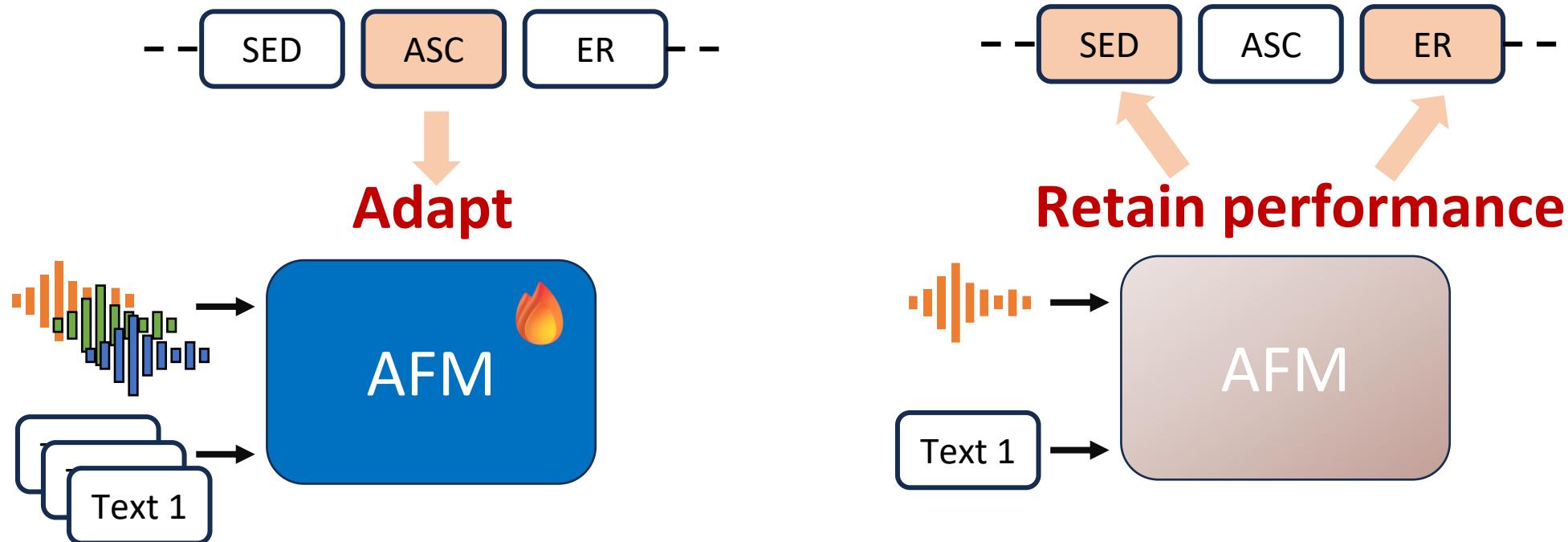
# Results: test-time adapt vs zero-shot baseline

	Average ↑	Sound Event Classification ↑			Vocal Sound Classification ↑	Surveillance Sound Classif.↑	Acoustic Scene Classification↑
Model	Average	ESC50	US8K	DCASE17 Task 4	Vocal Sound	SESA	TUT 2017
Zero-Shot	62.93	93.90	82.30	46.60	79.97	64.95	53.80
One audio	64.94	93.35	<b>85.26</b>	50.96	82.14	73.30	54.19
Five audio	<b>65.92</b>	<b>95.05</b>	85.21	<b>52.30</b>	<b>82.40</b>	<b>74.35</b>	<b>54.38</b>

	Music Classification ↑		Instrument Classification ↑		Speech Emotion Classification↑	
Model	GTZAN Music	GTZAN Speech	Beijing Opera	NS Instr. family	CRE MA-D	RAV DESS
Zero-Shot	99.20	58.40	46.60	68.00	30.00	31.54
One audio	99.21	61.00	47.45	68.28	29.92	<b>34.25</b>
Five audio	<b>100.0</b>	<b>63.20</b>	<b>50.42</b>	<b>69.23</b>	<b>31.27</b>	33.19

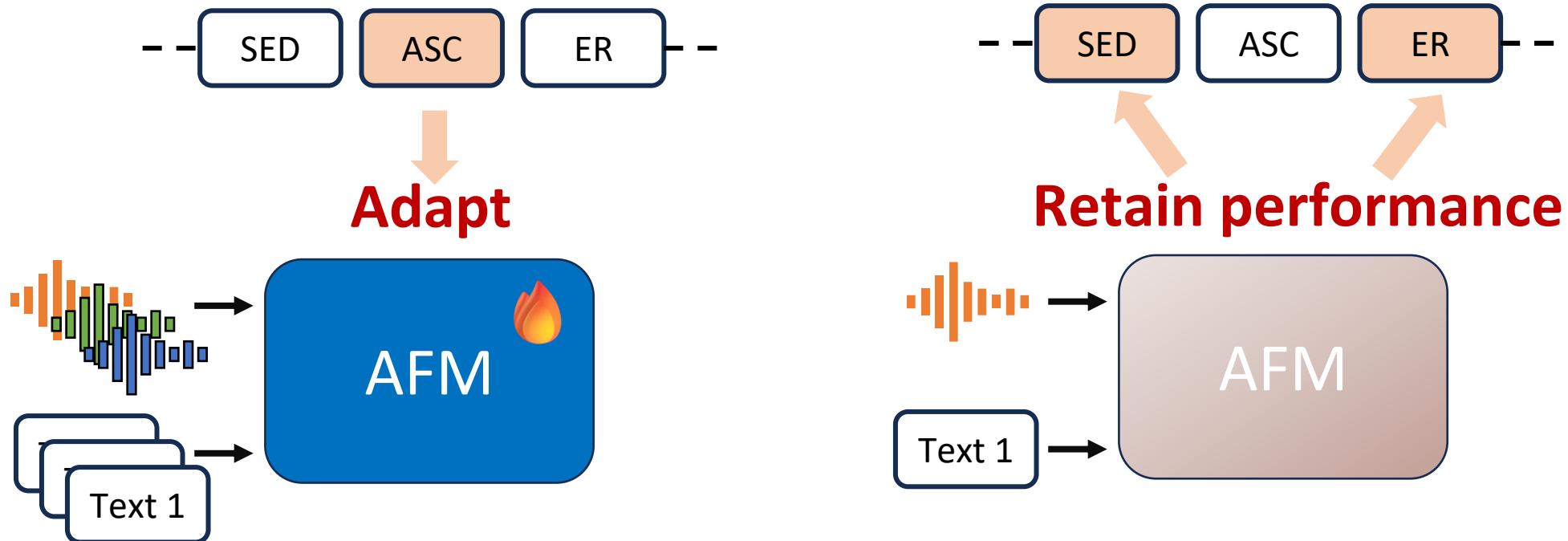
# Retaining generalization

One key property of Audio Foundation Models is **generalization**

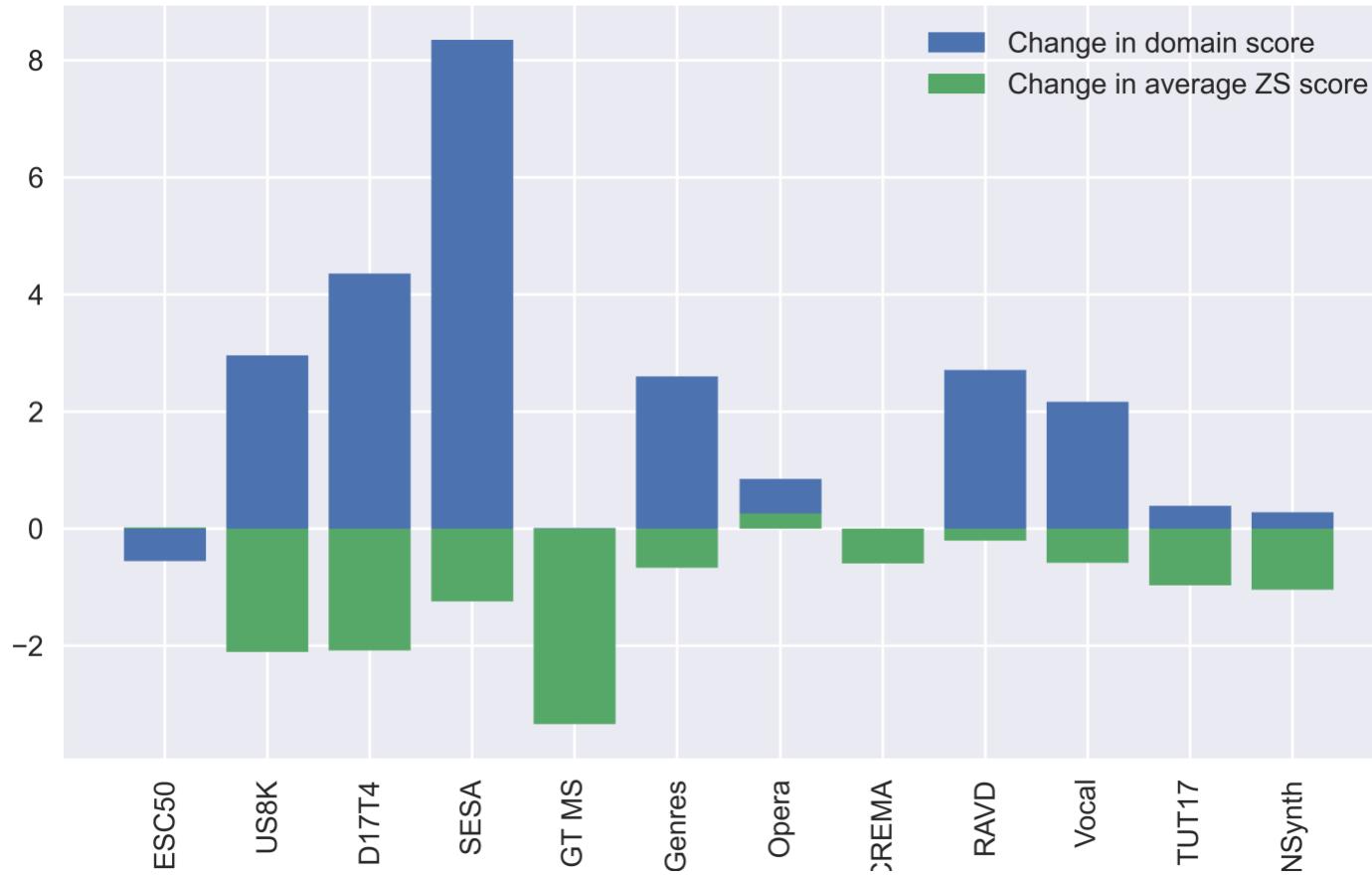


# To verify generalization

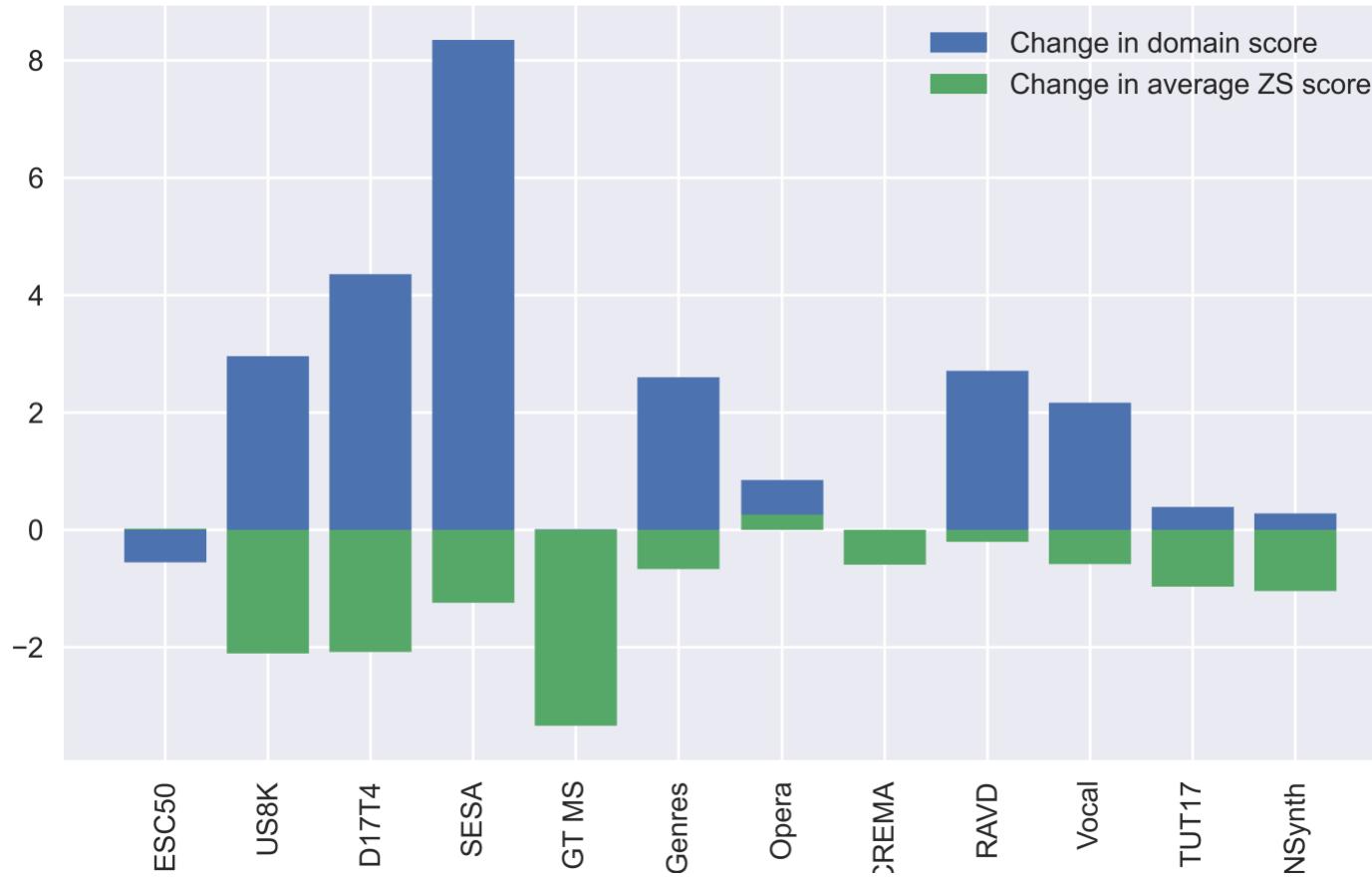
Adapt model on task 1 and then check its average performance on N-1 tasks



# Results: generalization



# Results: generalization



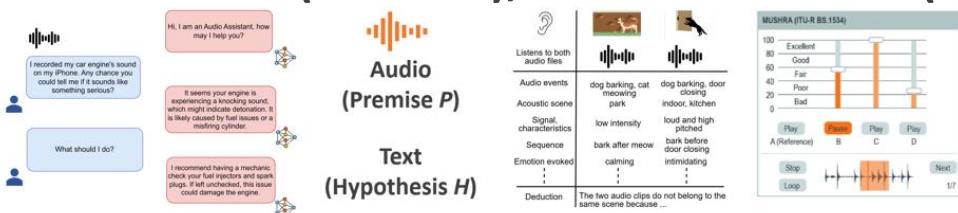
**3% increase for a drop of 1% in generalization**

# Thesis contributions

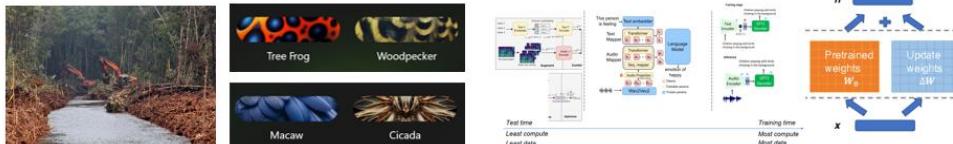
- Effective learning and pretraining methods  
**CLAP (ICASSP 23, 24), Pengi (NeurIPS 23), Mellow (under review)**



- Methods or benchmarks to analyze new abilities  
**Audio Entail (AAAI 25), Audio Difference (ICLR 25), PAM (INTERSPEECH 24)**



- Efficient adaptation methods to enable customization  
**Test-time (ICASSP 24), Few-shot (INTERSPEECH 25), Missing modality (ICASSP 25)**



# A short interlude - what's next?

# A short interlude - what's next?

- Symbolic reasoning in Audio Foundation Models
- Generalized Audio Foundation Models
- Reasoning vs approximate retrieval
- Beyond language as a mode of supervision

# Publications

- ❑ B. Elizalde, **S. Deshmukh**, M. A. Ismail and H. Wang, "CLAP Learning Audio Concepts from Natural Language Supervision," ICASSP 2023
- ❑ **S. Deshmukh**, B. Elizalde and H. Wang, "Audio Retrieval with WavText5K and CLAP Training" INTERSPEECH 2023
- ❑ **S. Deshmukh**, B. Elizalde, R. Singh and H. Wang, "Pengi: Audio Language Model for Audio Tasks" NeurIPS 2023
- ❑ B. Elizalde\*, **S. Deshmukh\***, and H. Wang, "Natural Language Supervision for General Purpose Audio Representation" ICASSP 2024
- ❑ **S. Deshmukh**, B. Elizalde, D. Emmanouilidou, B. Raj, R. Singh, and H. Wang "Training Audio Captioning Models without Audio" ICASSP 2024
- ❑ **S. Deshmukh**, R Singh and B Raj, "Domain Adaptation for Contrastive Audio-Language Models" INTERSPEECH 2024
- ❑ H Bukhari, **S Deshmukh**, H Dhamyal, B Raj and R Singh, "SELM: Enhancing Speech Emotion Recognition for Out-of-Domain Scenarios" INTERSPEECH 2024
- ❑ **S Deshmukh**, D Alharthi, B Elizalde, H Gamper, R Singh and B Raj "PAM: Prompting Audio-Language Model for Audio Quality Assessment" INTERSPEECH 2024
- ❑ **S Deshmukh**, S Han, H Bukhari, B Elizalde, H Gamper, R Singh and B Raj "Audio Entailment: Assessing Deductive Reasoning for Audio Understanding" AAAI 2025 Oral
- ❑ S Dixit, **S. Deshmukh**, B Raj, "MACE: Leveraging Audio for Evaluating Audio Captioning Systems," ICASSP 2025 Workshop
- ❑ **S Deshmukh**, S Han, R Singh and B Raj "ADIFF: Explaining audio difference using natural language" ICLR 2025 Spotlight
- ❑ **S Deshmukh**, S Dixit, R Singh and B Raj "Mellow: small audio-language model for reasoning" Arxiv preprint, under review

# Acknowledgements

Advisor: Bhiksha Raj, Rita Singh

Thesis committee: Bhiksha Raj (CMU), Rita Singh (CMU), Shinji Watanabe (CMU), Dan Ellis (Google DeepMind), Benjamin Elizalde (Apple)

Microsoft speech: Huaming Wang, Min Tang, Raymond Xia, Hemin Yang, Mahmoud Al Ismail, Daniel Thompkins

Microsoft research: Dimitra Emmanouilidou, Hannes Gamper, Sebastian Braun, Ivan Tashev

CMU collaborators: Shuo Han, Hazim Bukhari, Dareen Alharthi, Satvik Dixit

Robust-MLSP: Ahmed Shah, Roshan Sharma, Hira Dhamyal, Mark Lindsay, Raphael Oliver, Tyler Vuong, Muqiao Yang, Yandong Wen, Wayne Zhao, Ankit Shah, Joseph Konan, Xiang Li, Hao Chen, Muqiao Yang, Syed Abdul Hannan, Massa Baali, Kate Shapovalenko, Abdul Waheed, Abdulhamid Aldoobi, Sarthak Bisht, Karan Singh, Shravanth Srinivas, Kelsey Harvey, Kashu Yamazaki, Sunghwan Baek, ...

**and many more for inspiration**

# Acknowledgements



# Thesis summary and thank you for listening

**This thesis investigates two central questions –**

- What are effective learning or pretraining methods to enable the scaling of audio models?
- Once pretrained, how can the new abilities be analyzed, evaluated, and subsequently improved?

- **Effective learning and pretraining methods**

- CLAP (ICASSP 23, 24), Pungi (NeurIPS 23), Mellow (under review)

- **Methods or benchmarks to analyze new abilities**

- Audio Entail (AAAI 25 Oral), Audio Difference (ICLR 25 Spotlight), PAM (INTERSPEECH 24)

- **Efficient adaptation methods to enable customization**

- Test-time (ICASSP 24), Few-shot (INTERSPEECH 25), Missing modality (ICASSP 25)