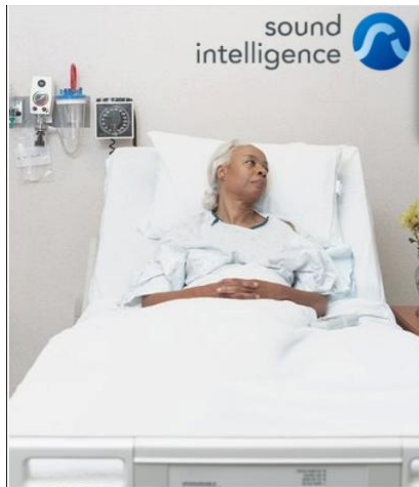


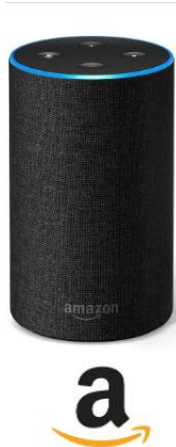
# Audio Foundation Models

Soham Deshmukh  
ECE department  
Carnegie Mellon University

# Audio Understanding has multiple applications



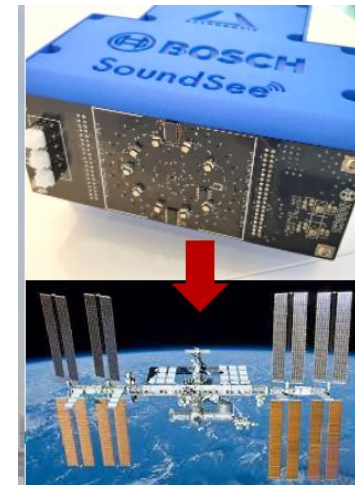
Healthcare



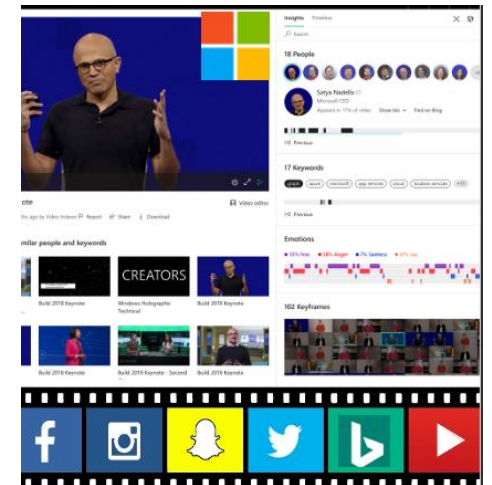
Safety



Noise Monitoring



Predictive  
Maintenance

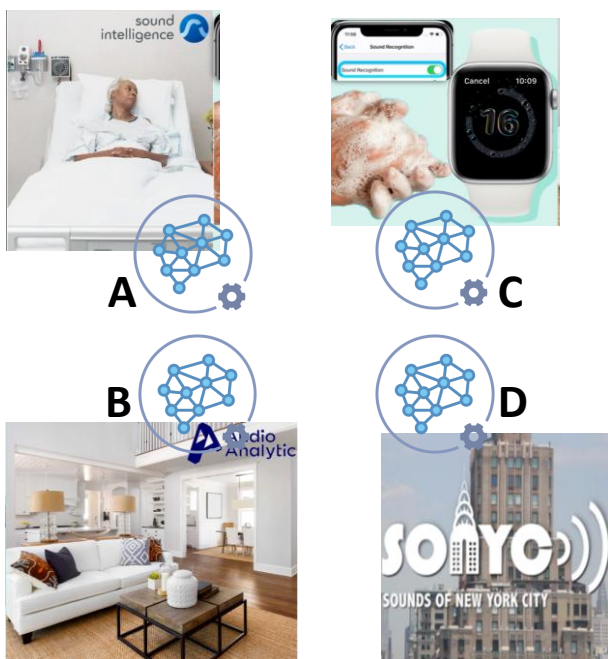


Content  
Retrieval

# ML models built for each task and domain



# Task-specific to Foundation Models



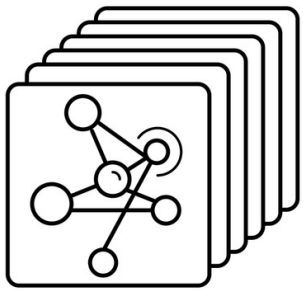
Towards foundation  
model paradigm



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

# Audio Foundation Model

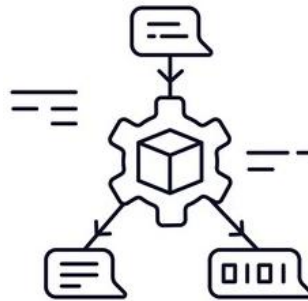
Multi-purpose ML model pre-trained on extensive audio datasets using self-supervised learning (SSL)



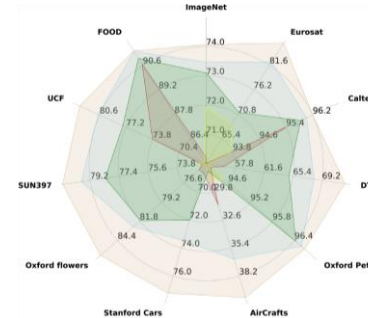
Pretraining on large datasets



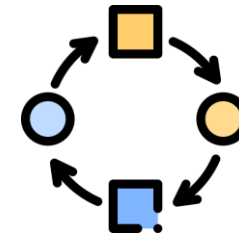
Self-Supervised Learning



Multi-modal capabilities



Generalization and versatility

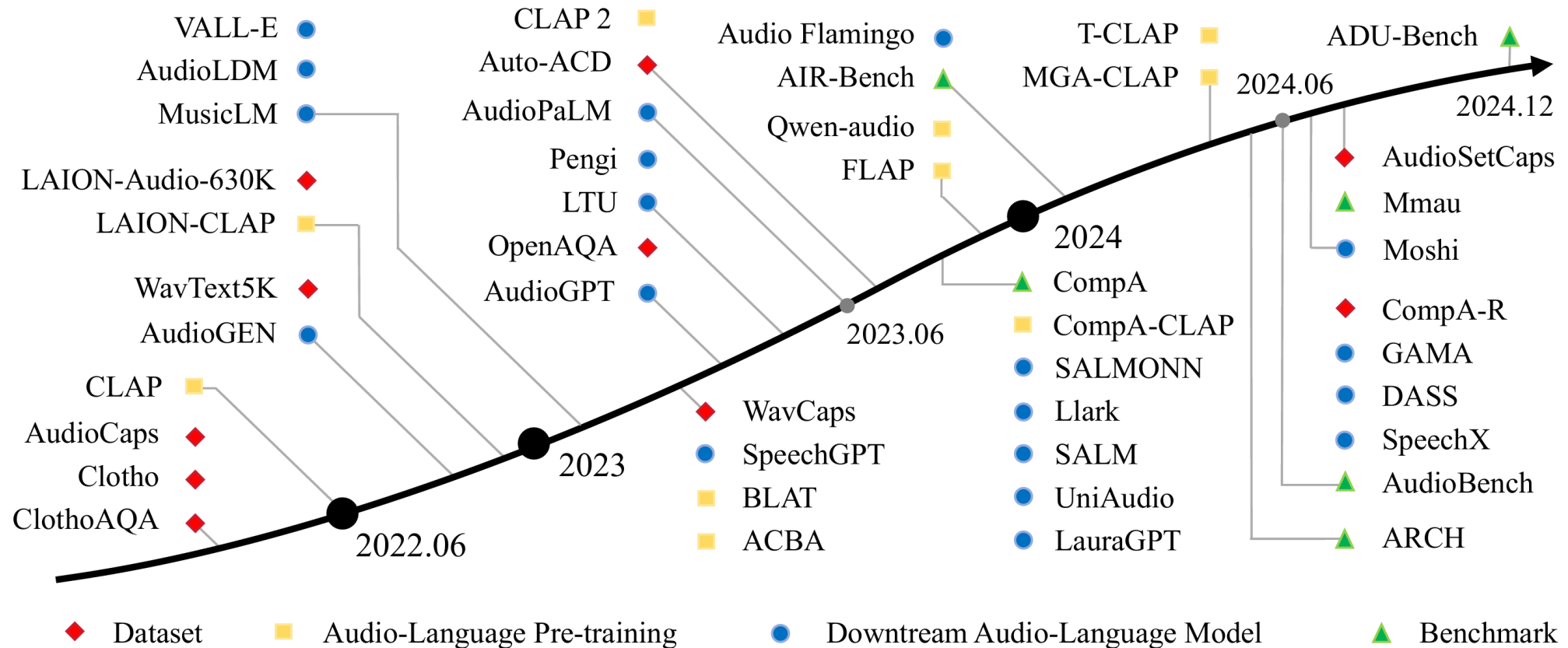


Adaptability



Emergent abilities

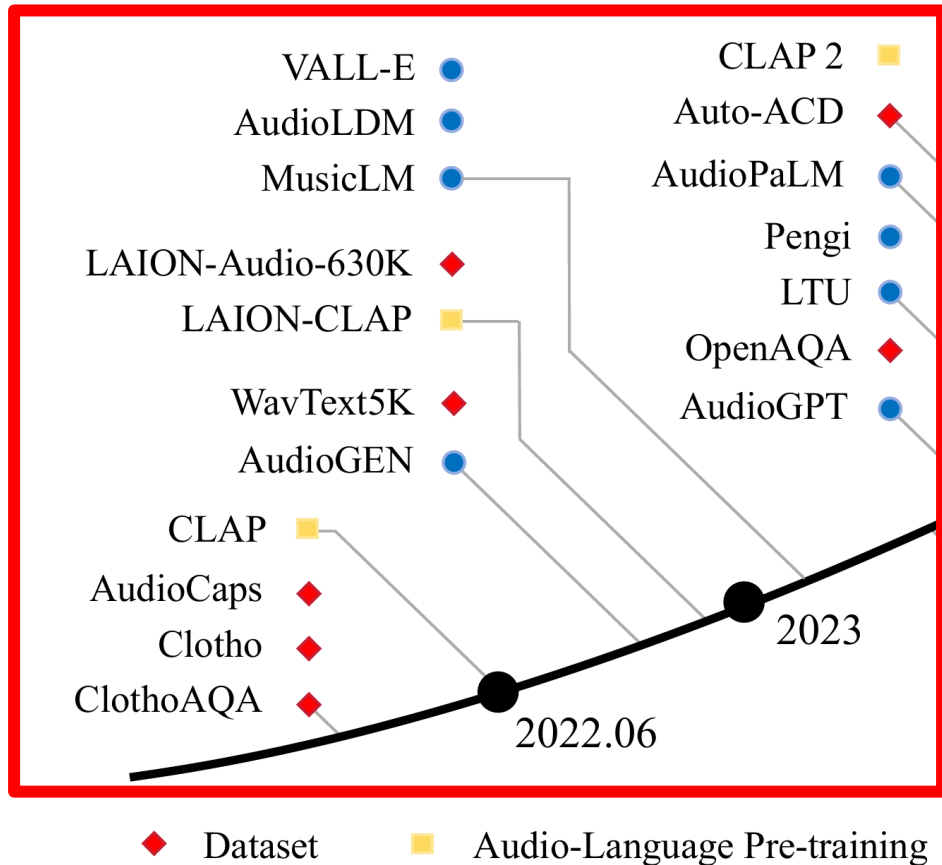
# Audio Foundation Model literature



Audio-Language Models for Audio-Centric Tasks: A survey:

<https://arxiv.org/abs/2501.15177>

# Audio Foundation Model literature

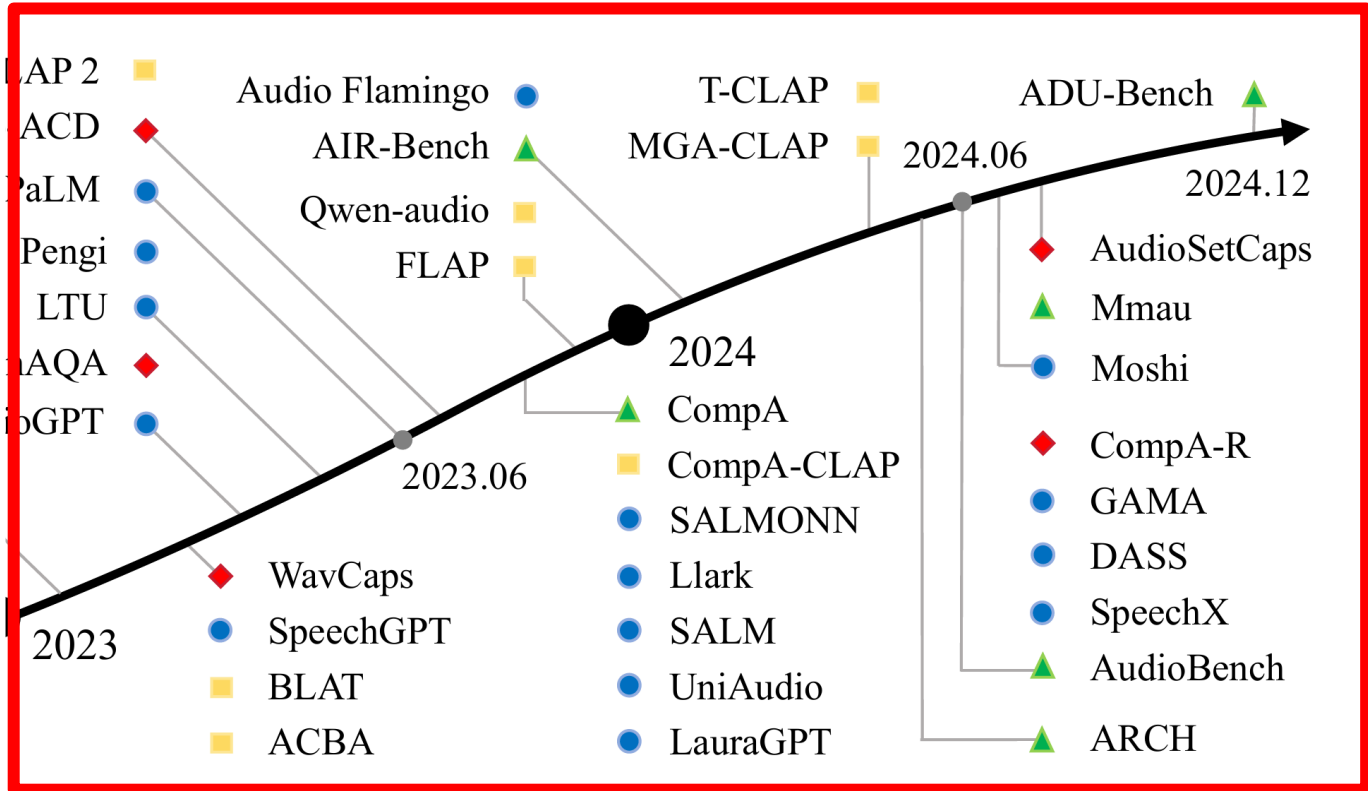


Contrastive pretraining for close-ended tasks like classification and retrieval

Audio-Language Models for Audio-Centric Tasks: A survey:  
<https://arxiv.org/abs/2501.15177>

# Audio Foundation Model literature

Generative pretraining  
for both close-ended  
and open-ended tasks



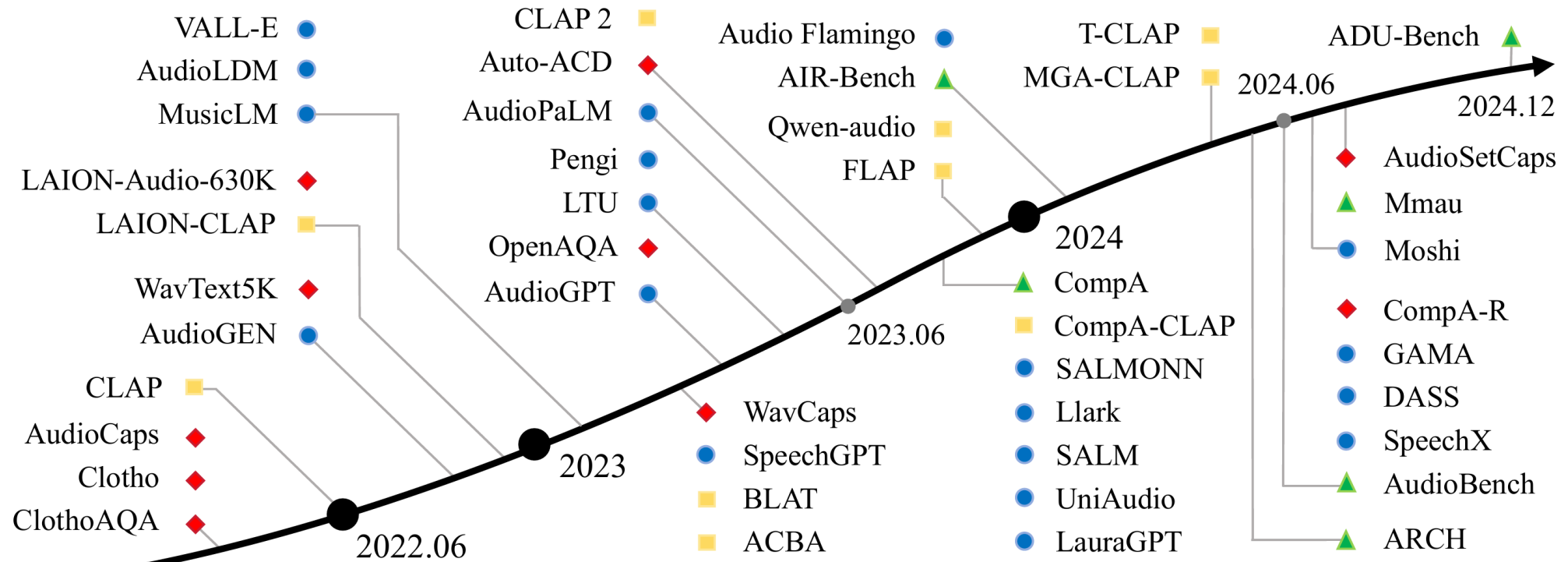
◆ Dataset      ■ Audio-Language Pre-training      ● Downstream Audio-Language Model      ▲ Benchmark

Audio-Language Models for Audio-Centric Tasks: A survey:

<https://arxiv.org/abs/2501.15177>



# Audio Foundation Model literature



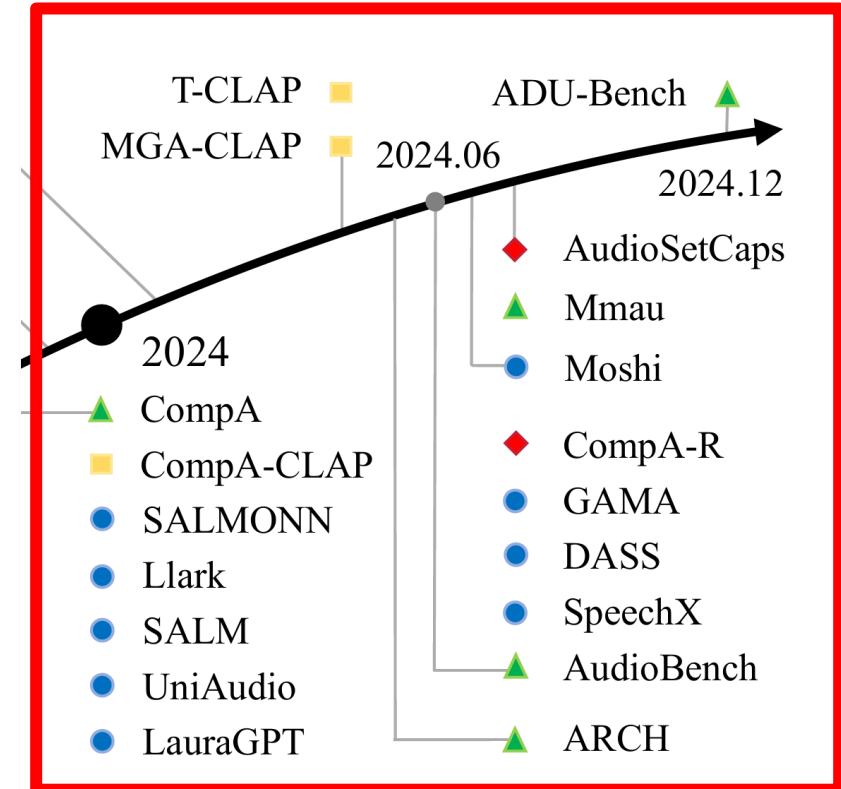
**“Simple” objective trained over millions of audio-text pairs**

Audio-Language Models for Audio-Centric Tasks: A survey:

<https://arxiv.org/abs/2501.15177>

# Audio Foundation Model literature

With scale, models start exhibiting new capabilities



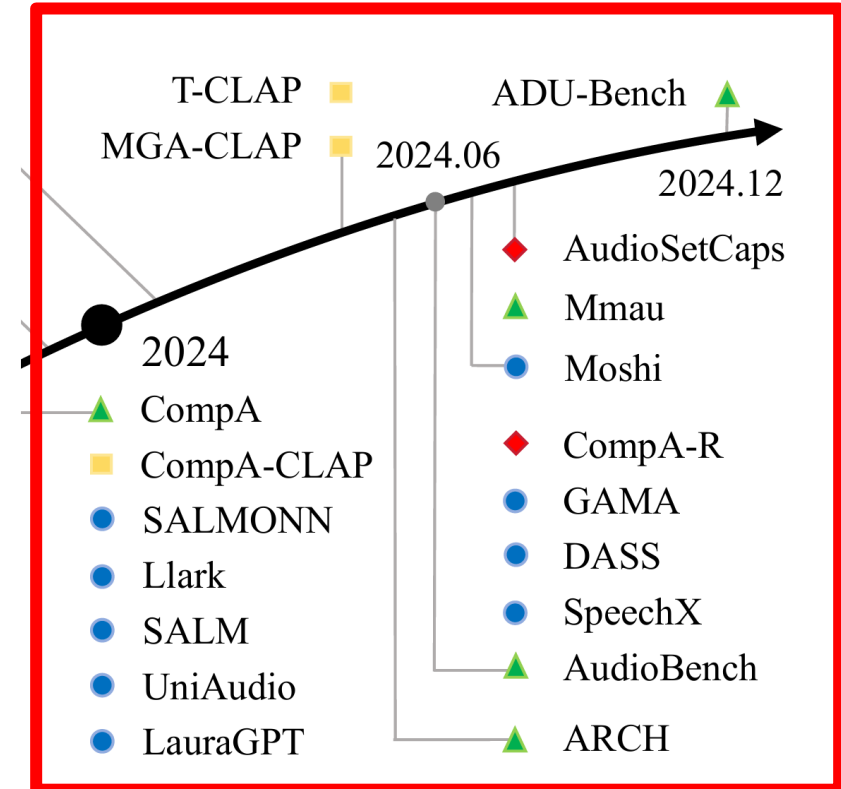
**“Simple” objective trained over millions of audio-text pairs**

Audio-Language Models for Audio-Centric Tasks: A survey:

<https://arxiv.org/abs/2501.15177>

# Audio Foundation Model literature

With scale, models start exhibiting new capabilities **including the ability to reason over both audio-text**

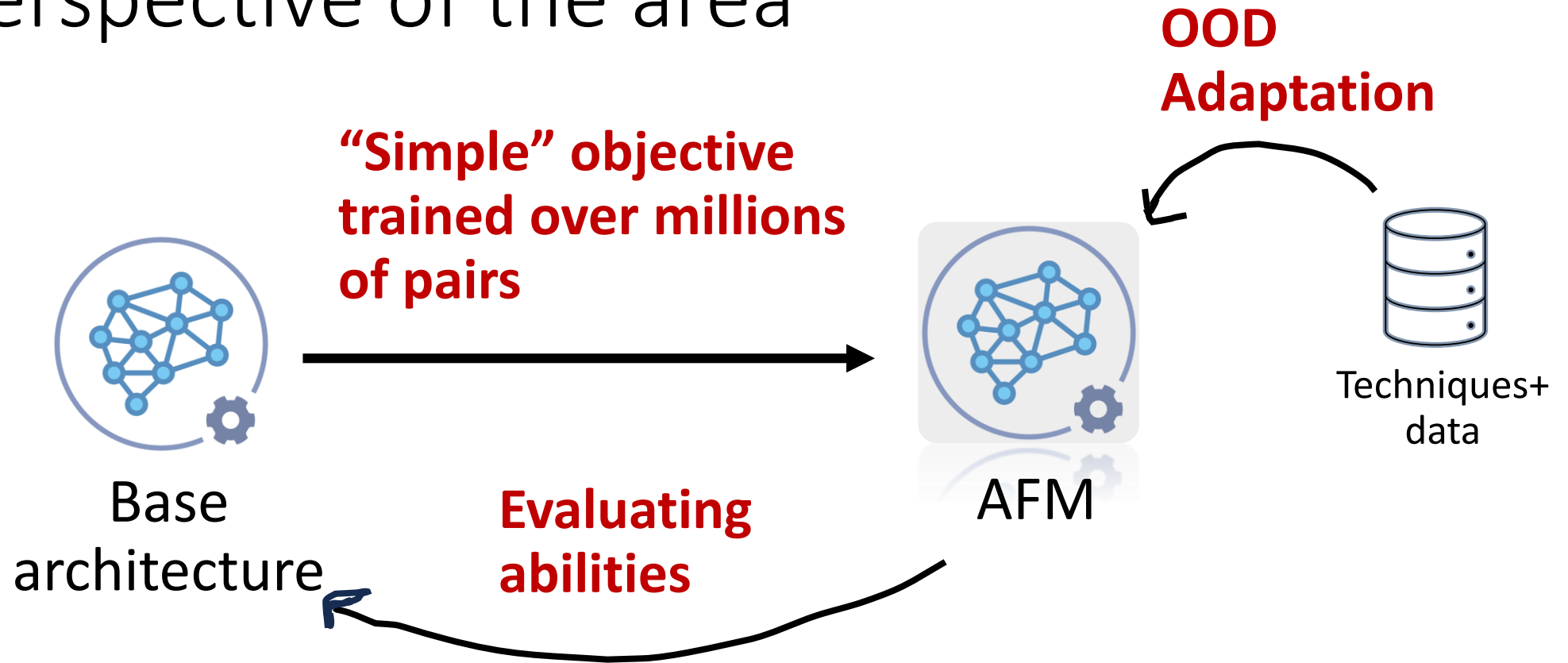


**“Simple” objective trained over millions of audio-text pairs**

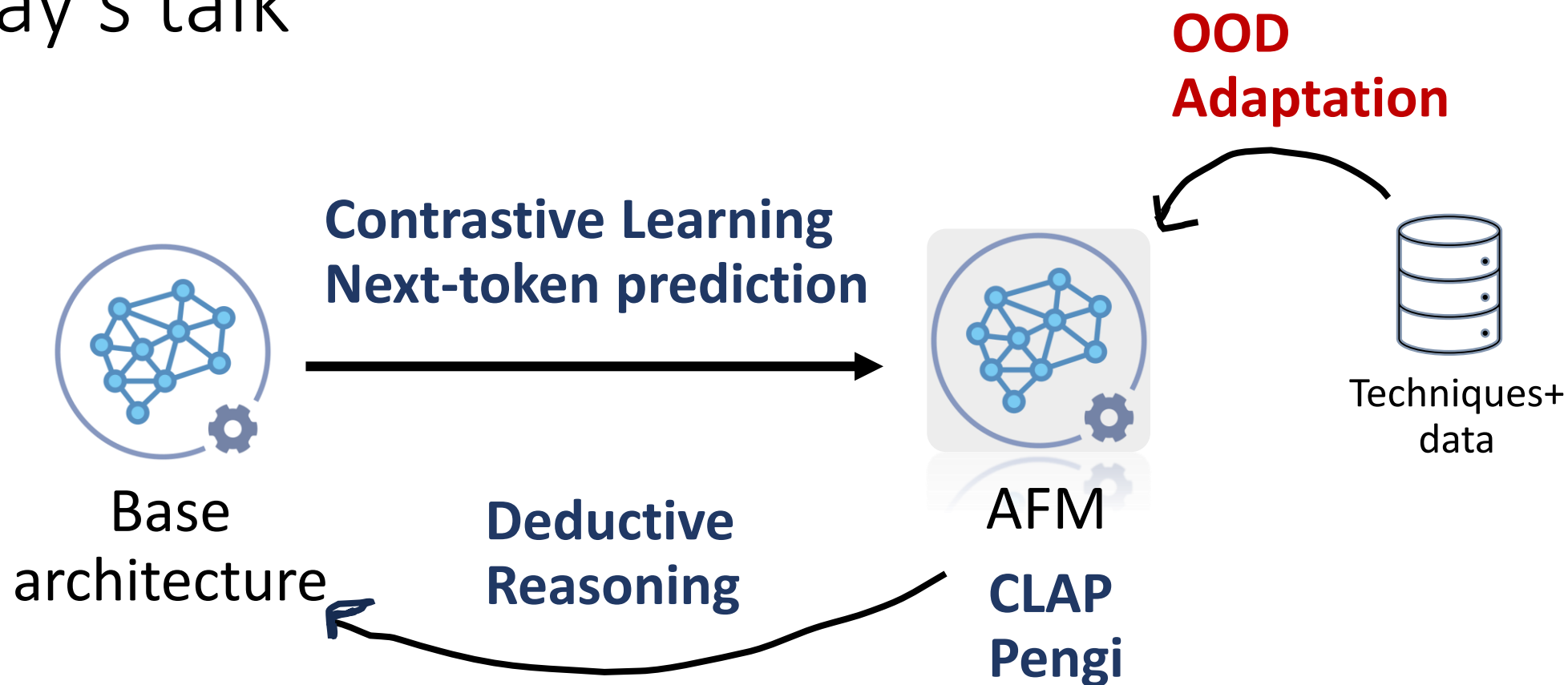
Audio-Language Models for Audio-Centric Tasks: A survey:

<https://arxiv.org/abs/2501.15177>

# A perspective of the area



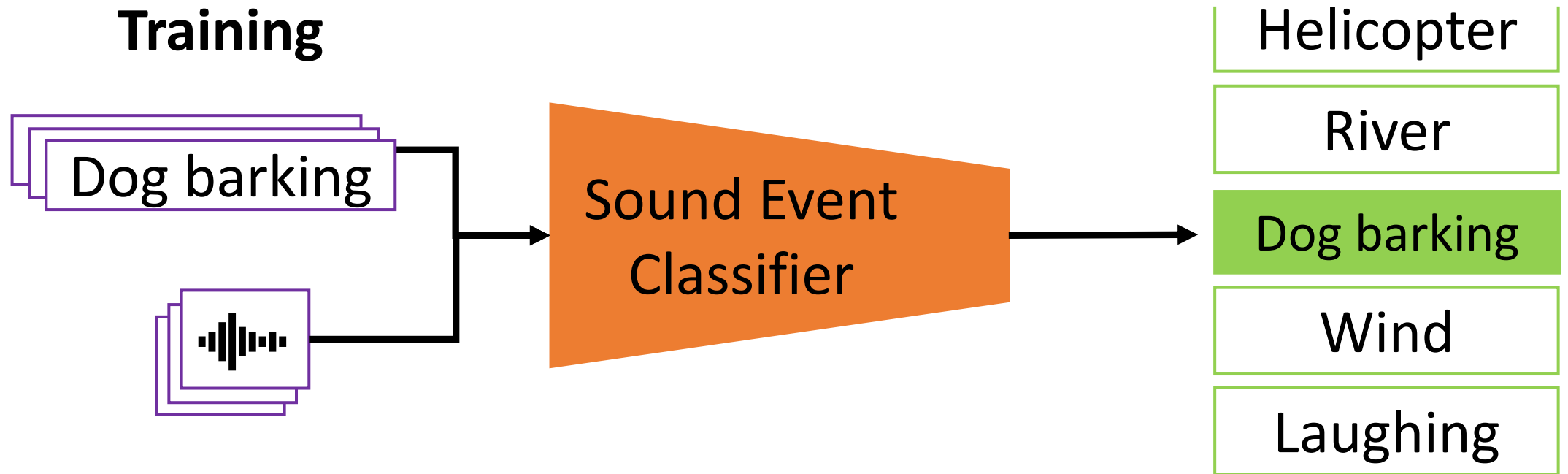
# Today's talk



# Talk outline

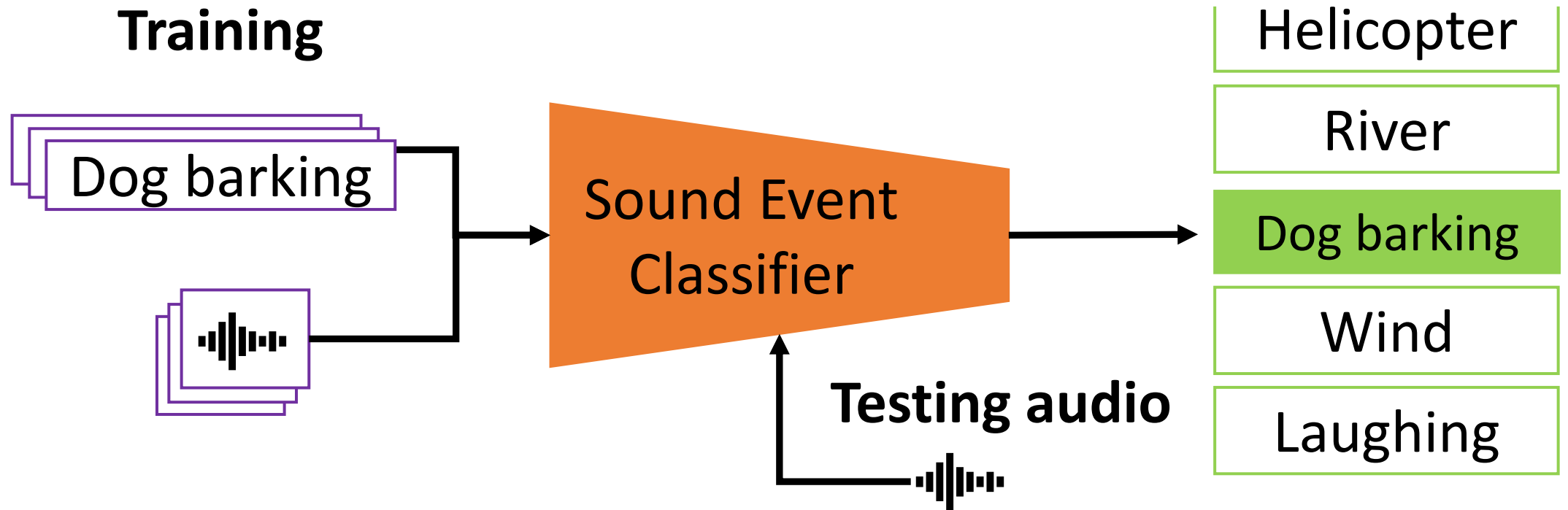
1. Contrastive Language-Audio Pretraining
2. Audio-conditioned next-token prediction
3. Deductive Reasoning

# A typical model training



Collecting annotated data and training model using supervised learning

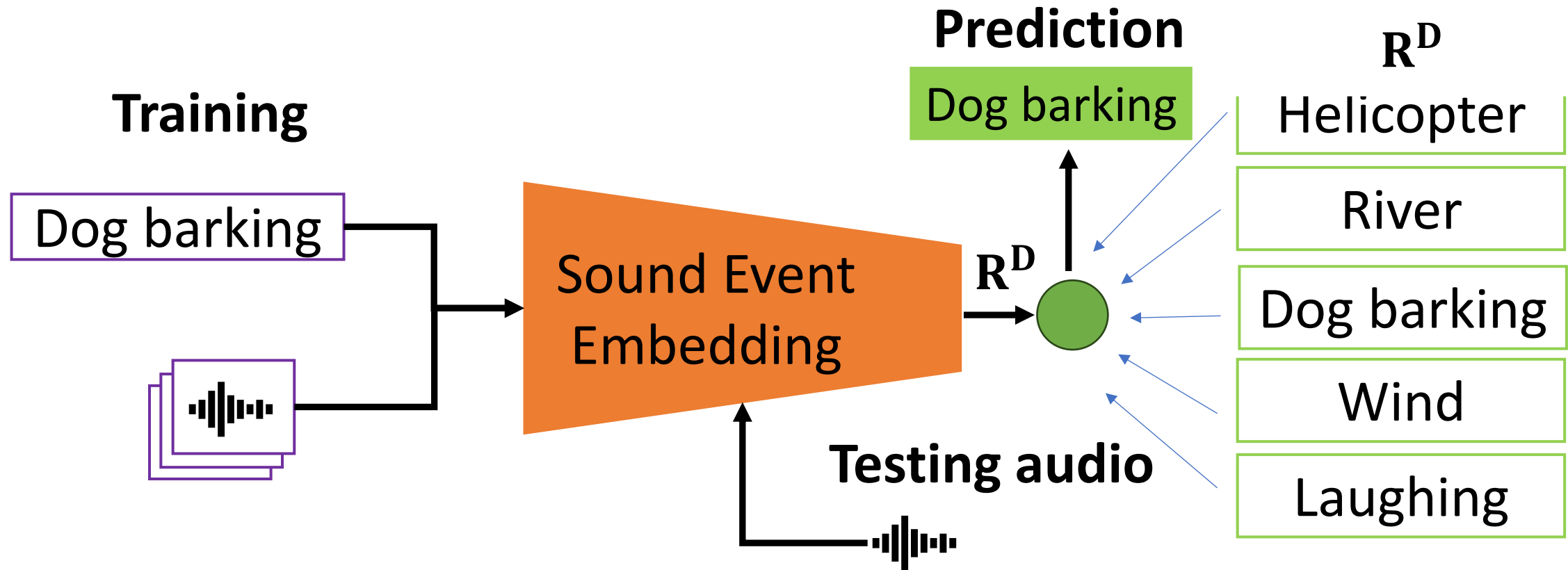
# Predicts predefined classes



During inference, the model predicts 1 out of N classes



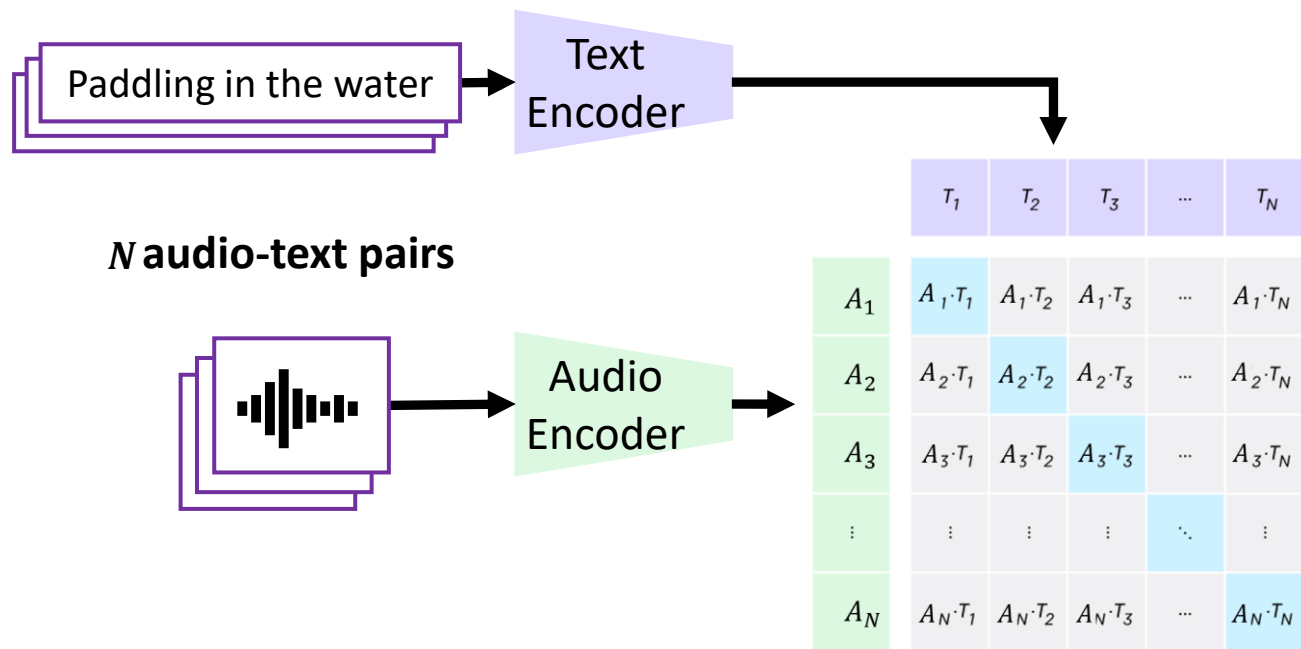
# Overcoming predefined classes



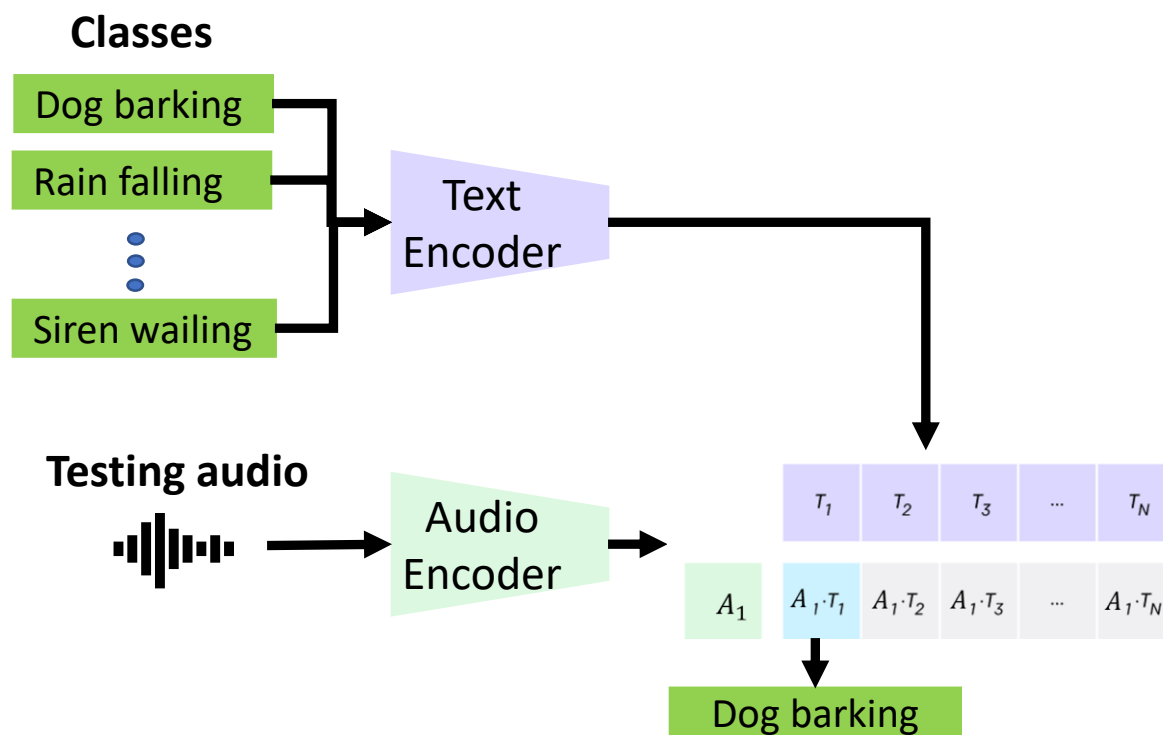
Move from classifier to embeddings and use similarity measure to determine prediction

# CLAP 🖐️ Contrastive Language-Audio Pretraining

## 1. Contrastive pretraining



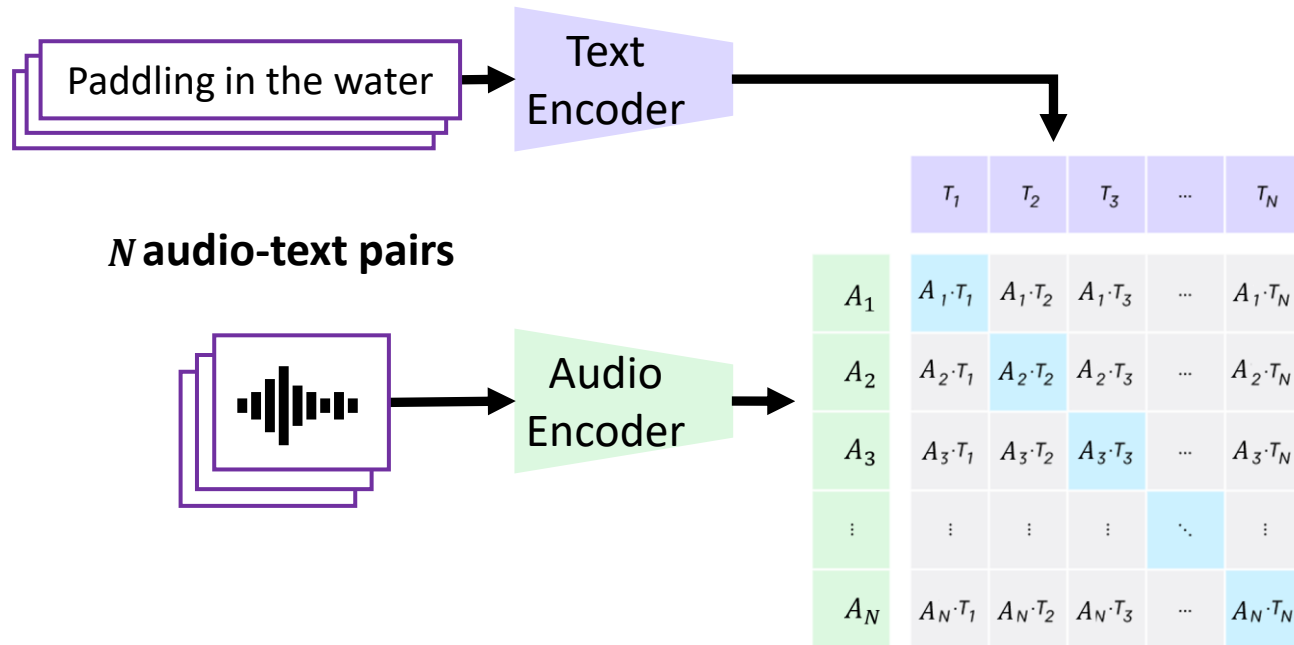
## 2. Zero-Shot classification



CLAP: Learning Audio Concepts From Natural Language  
Supervision <https://arxiv.org/abs/2206.04769>, ICASSP 2023

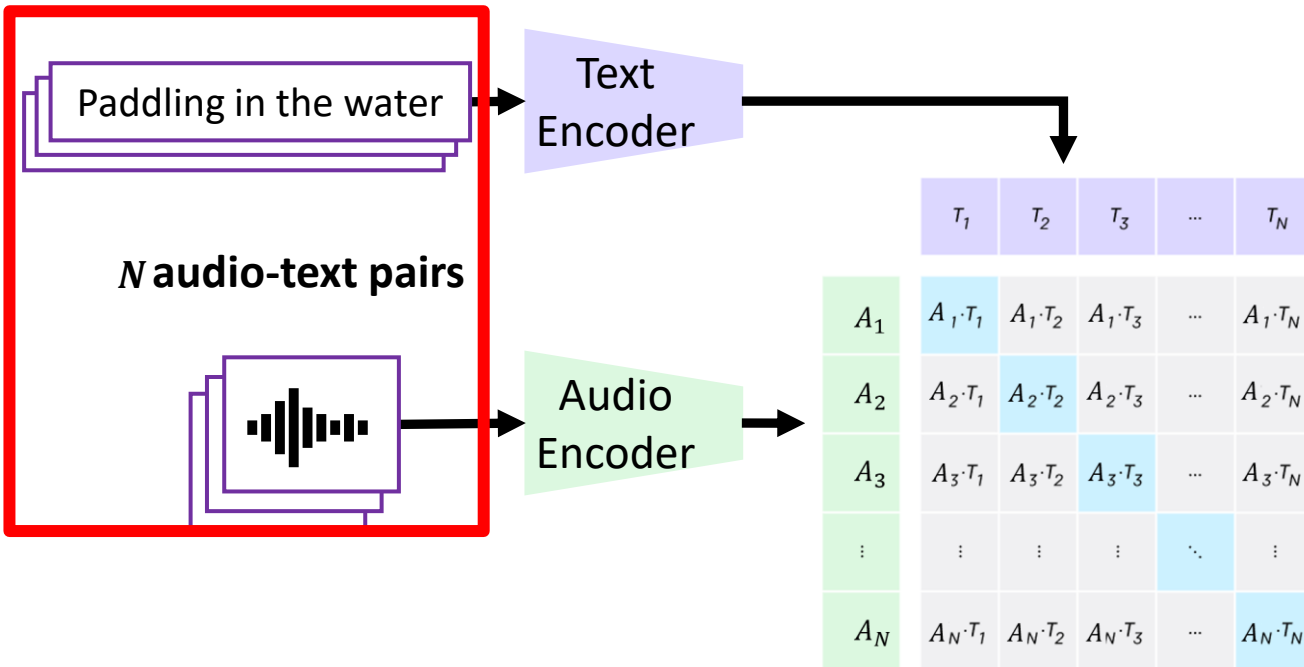
# Contrastive pretraining

## 1. Contrastive pretraining



# Contrastive pretraining

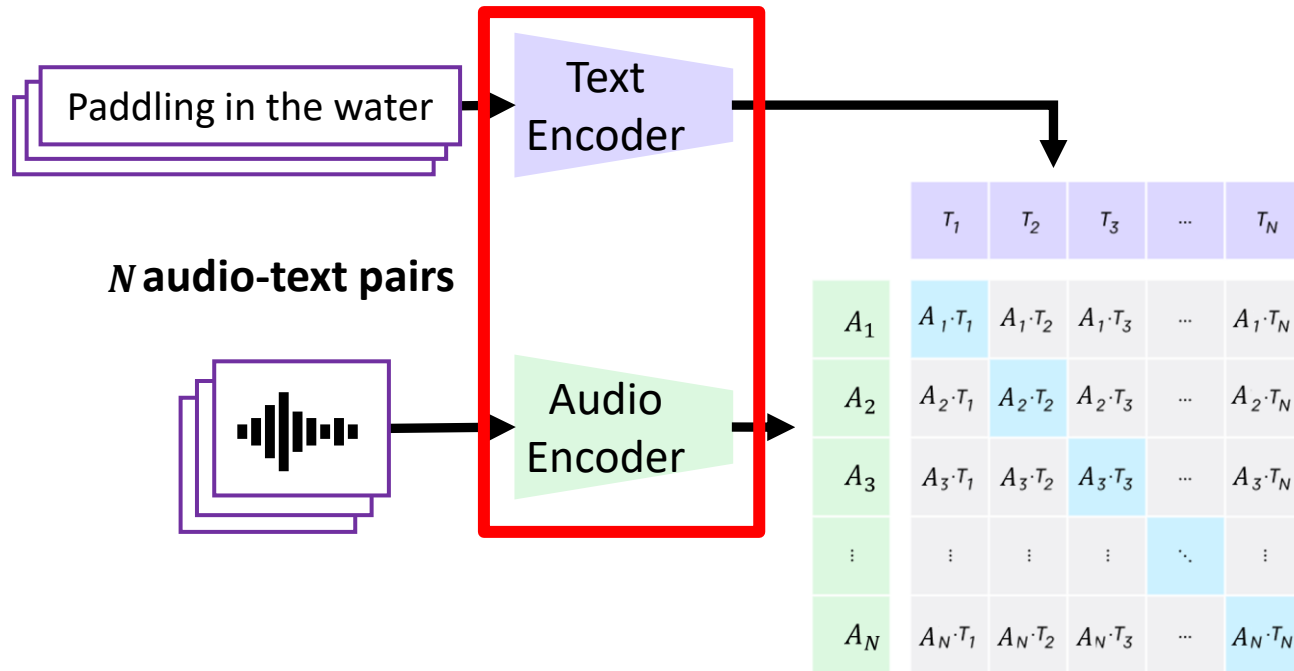
## 1. Contrastive pretraining



Training consists of batch of  $N$  audio-text pairs

# Contrastive pretraining

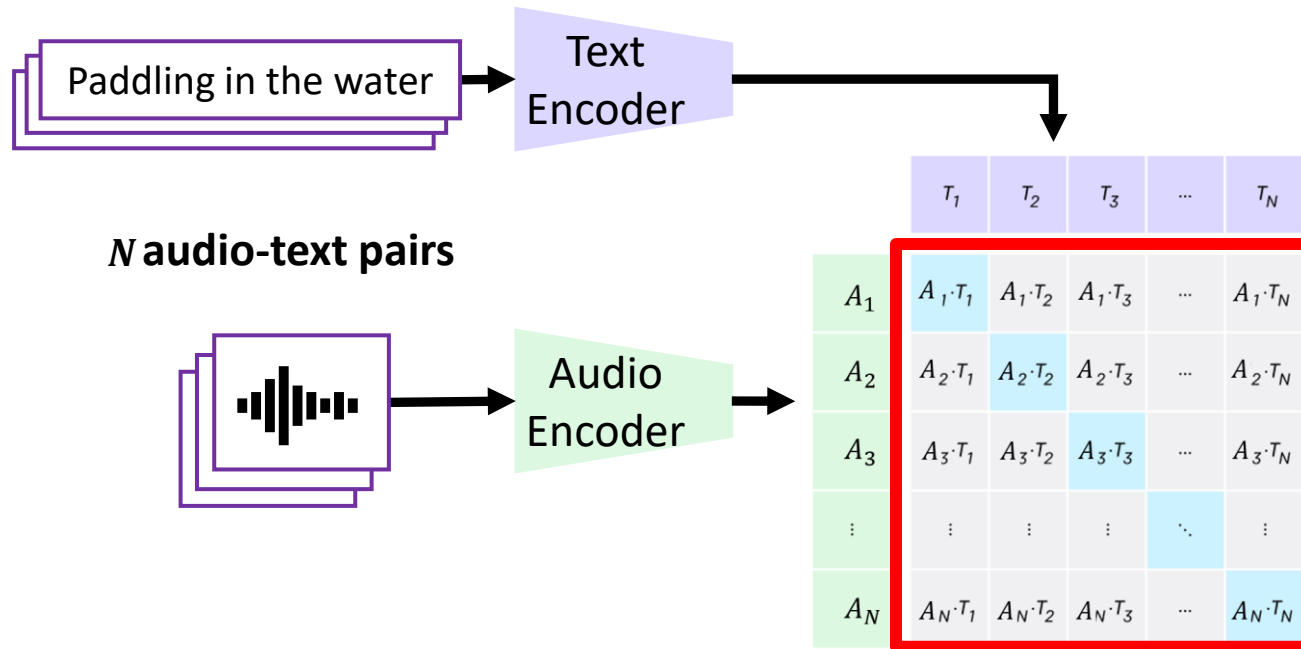
## 1. Contrastive pretraining



Independently encode audio-text pairs

# Contrastive pretraining

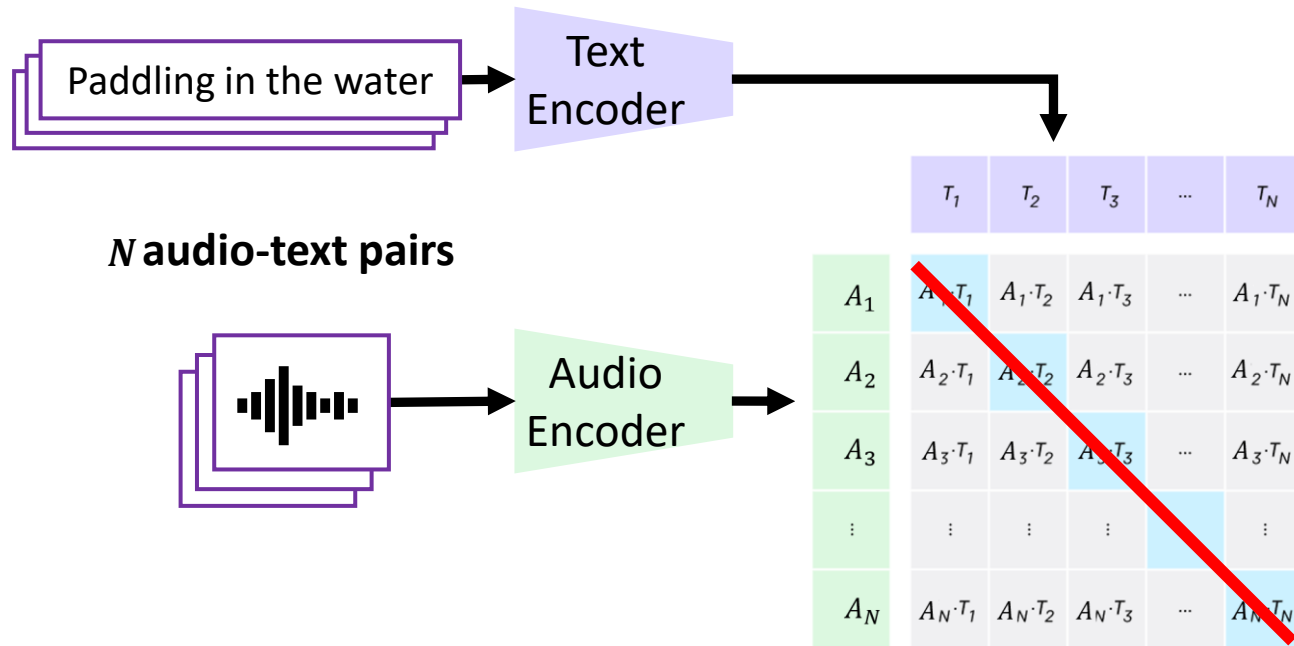
## 1. Contrastive pretraining



Compute dot product to form similarity matrix

# Contrastive pretraining

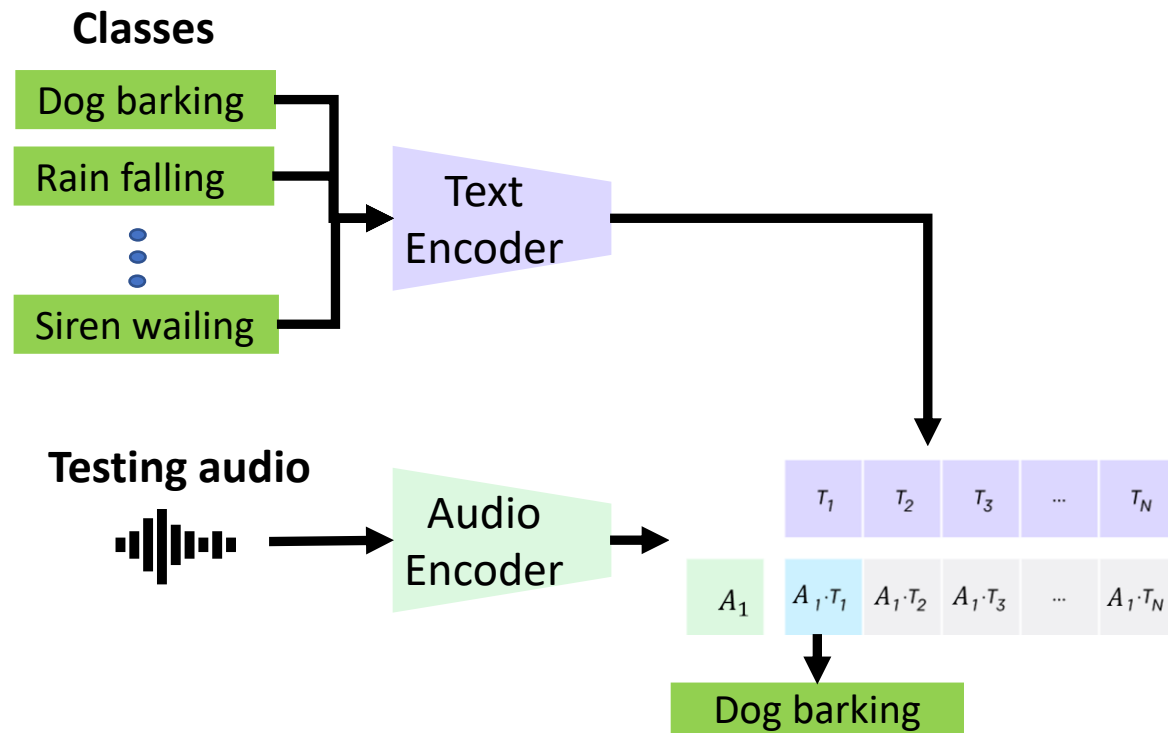
## 1. Contrastive pretraining



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

	Sound Event Classification					Music			
Model	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

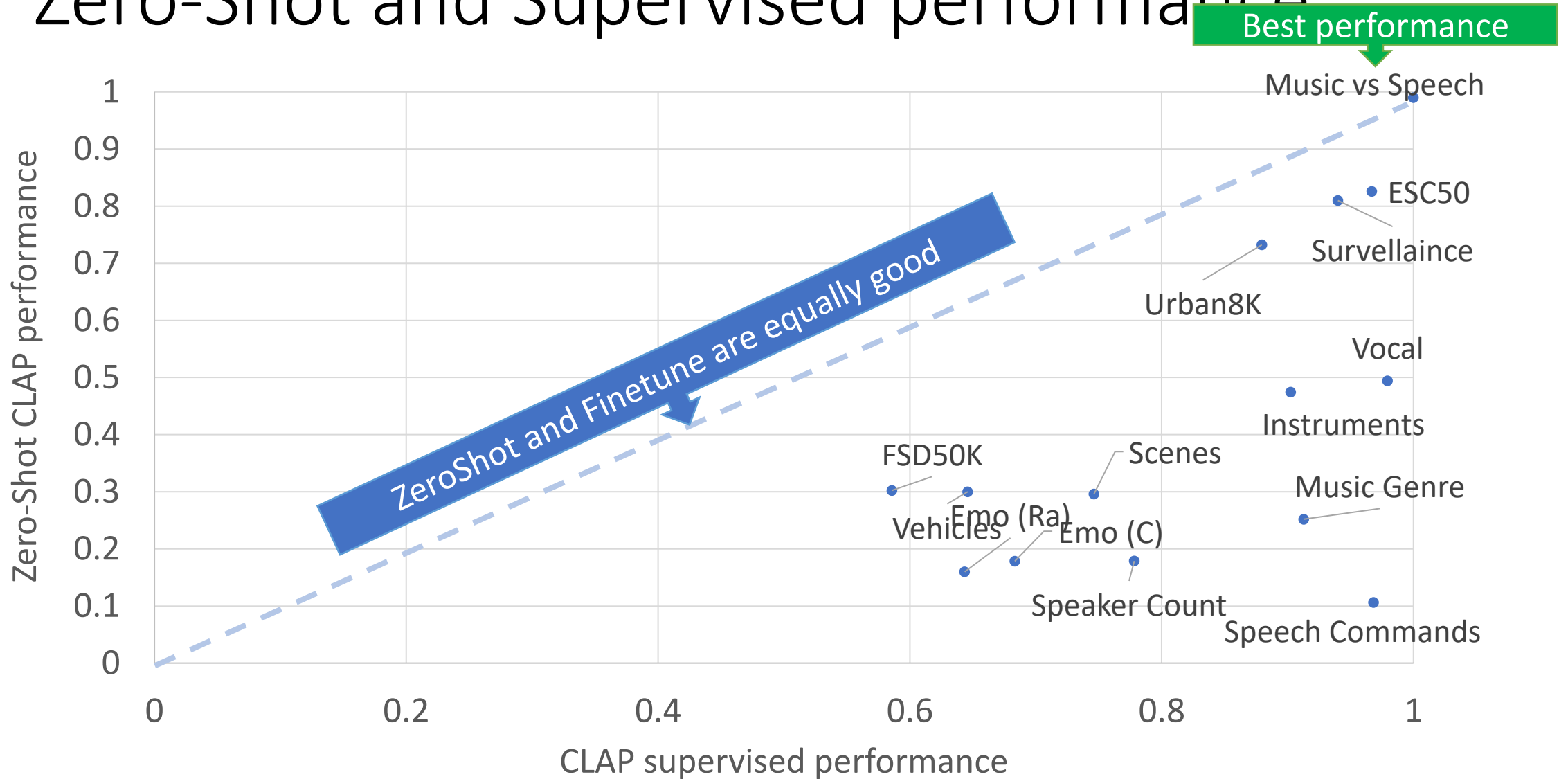
	Sound Event Classification					Music			
Model	ESC50	FSD50K	US8K	DCASE17 Task 4	AudioSet	Music Speech	Music Genres	Mri. Stroke	Mri. Tonic
Random	0.02	< 0.005	0.1	0.05	< 0.0018	0.5	0.1	0.1	0.1667
Benchmark (ZS)	0.6940[10]	0.0302[9]	0.6531[10]	-	-	-	-	-	-
CLAP(ZS)	<b>0.826</b>	<b>0.3024</b>	<b>0.7324</b>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

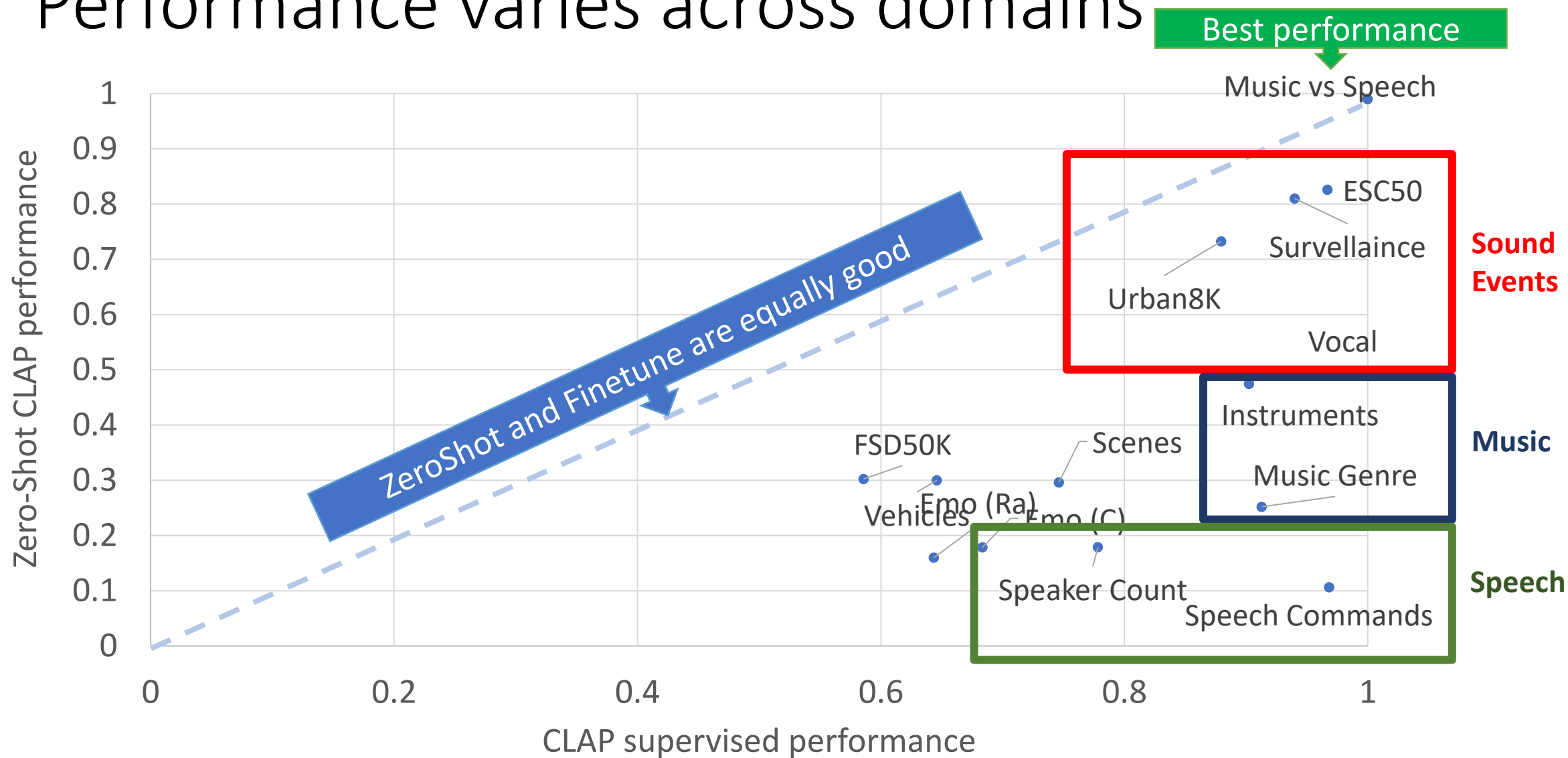
**Table 1.** CLAP (ZS) Zero-Shot outperforms the literature.

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

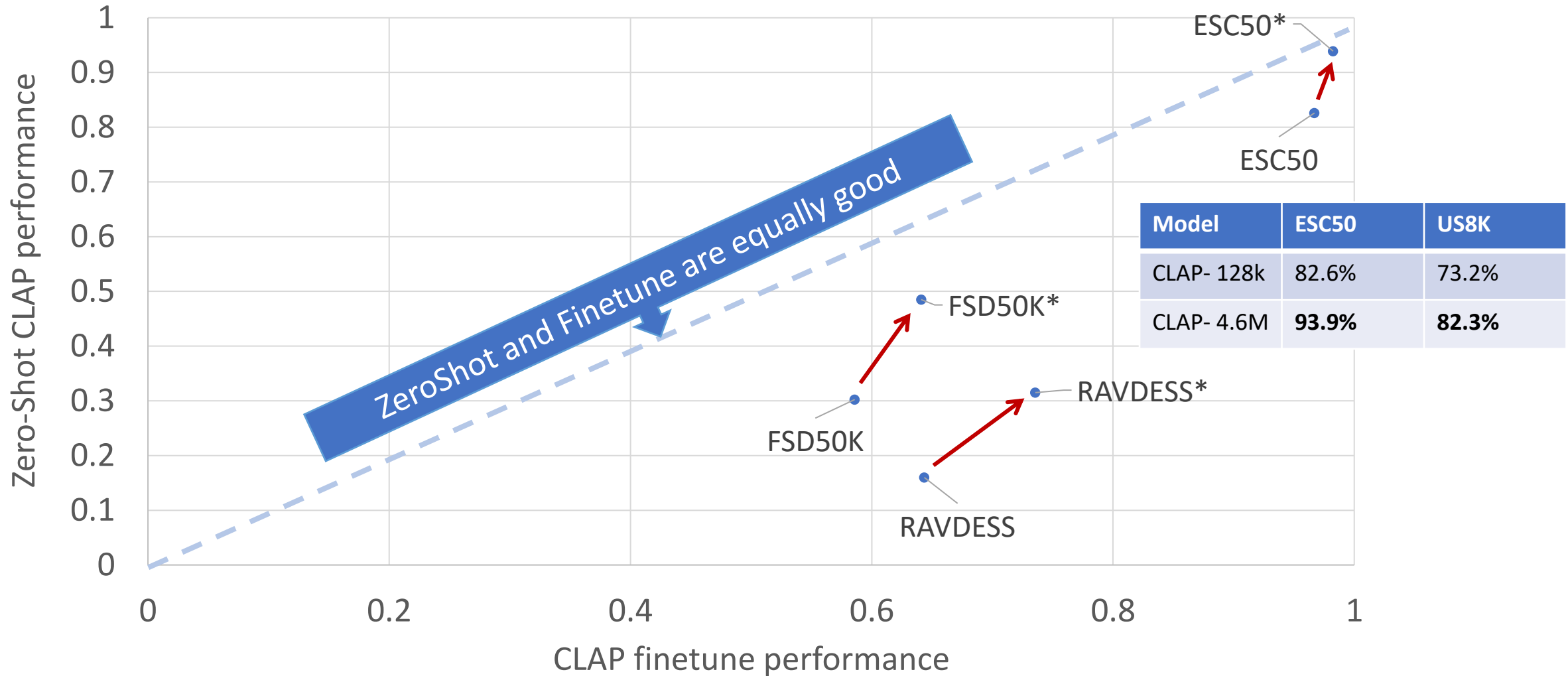
# Zero-Shot and Supervised performance



# 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?**

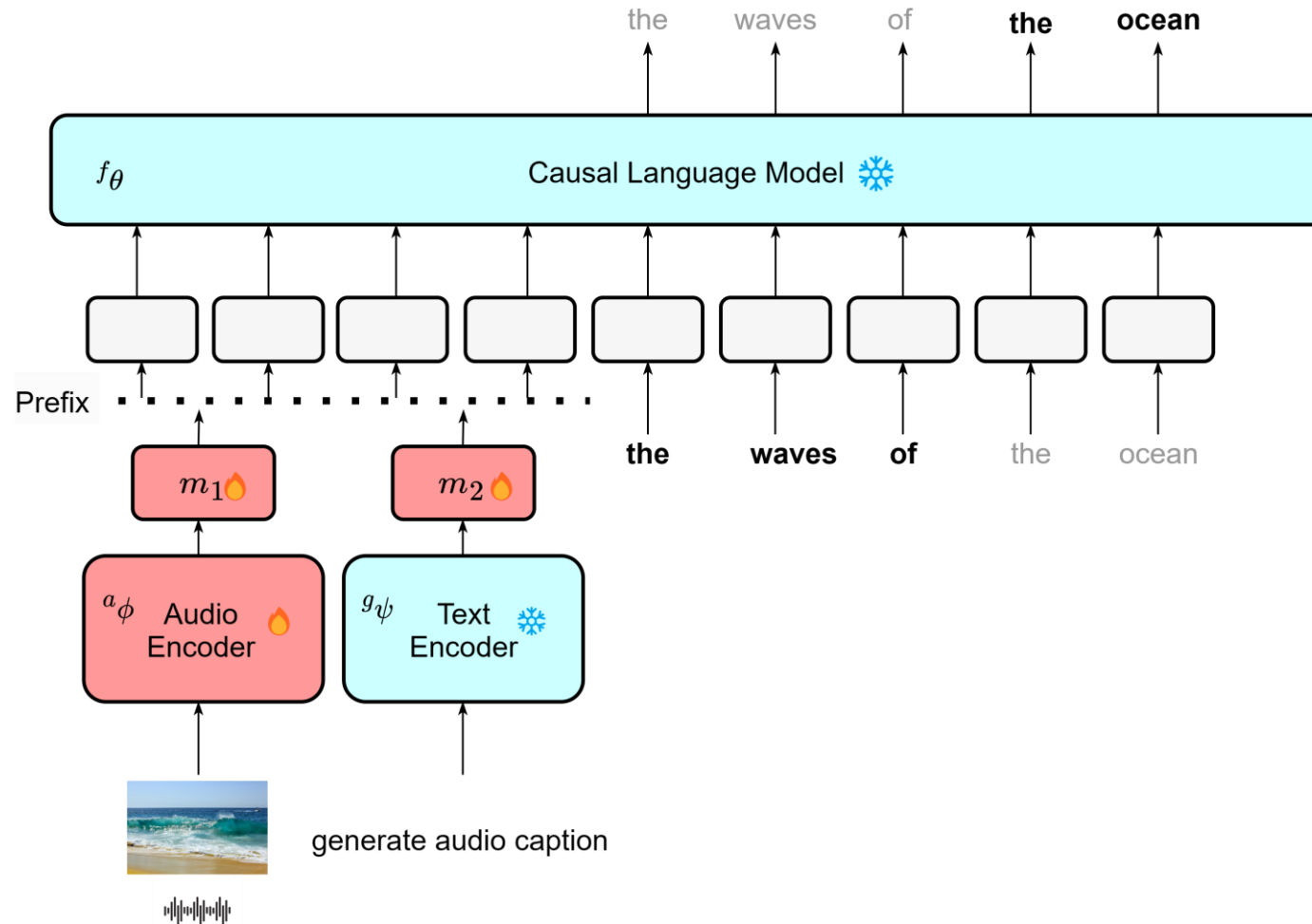
# Talk outline

1. Contrastive-Language Audio Pretraining
2. Audio-conditioned next-token prediction
3. Deductive Reasoning

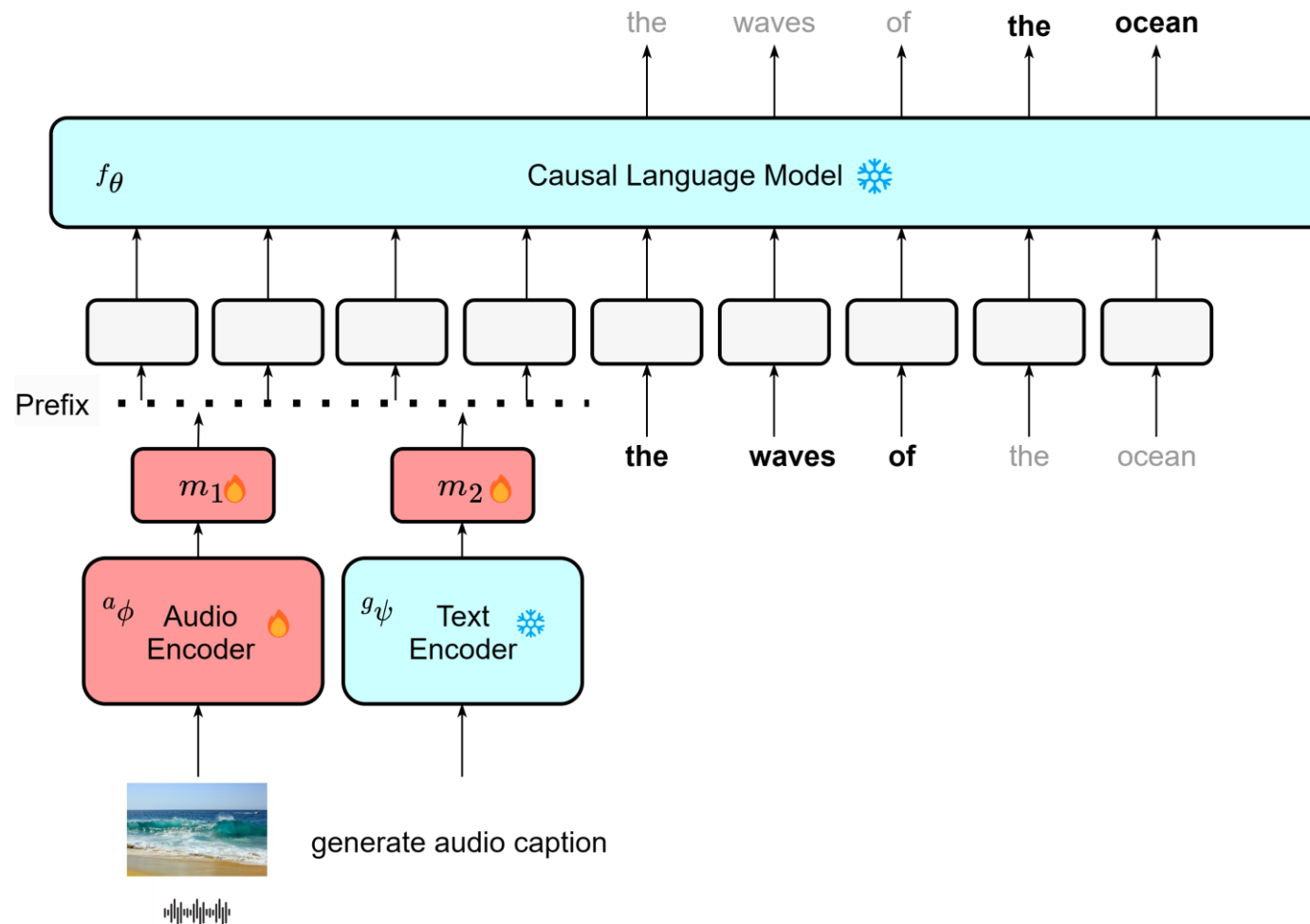
# Pengi: An Audio Language Model

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

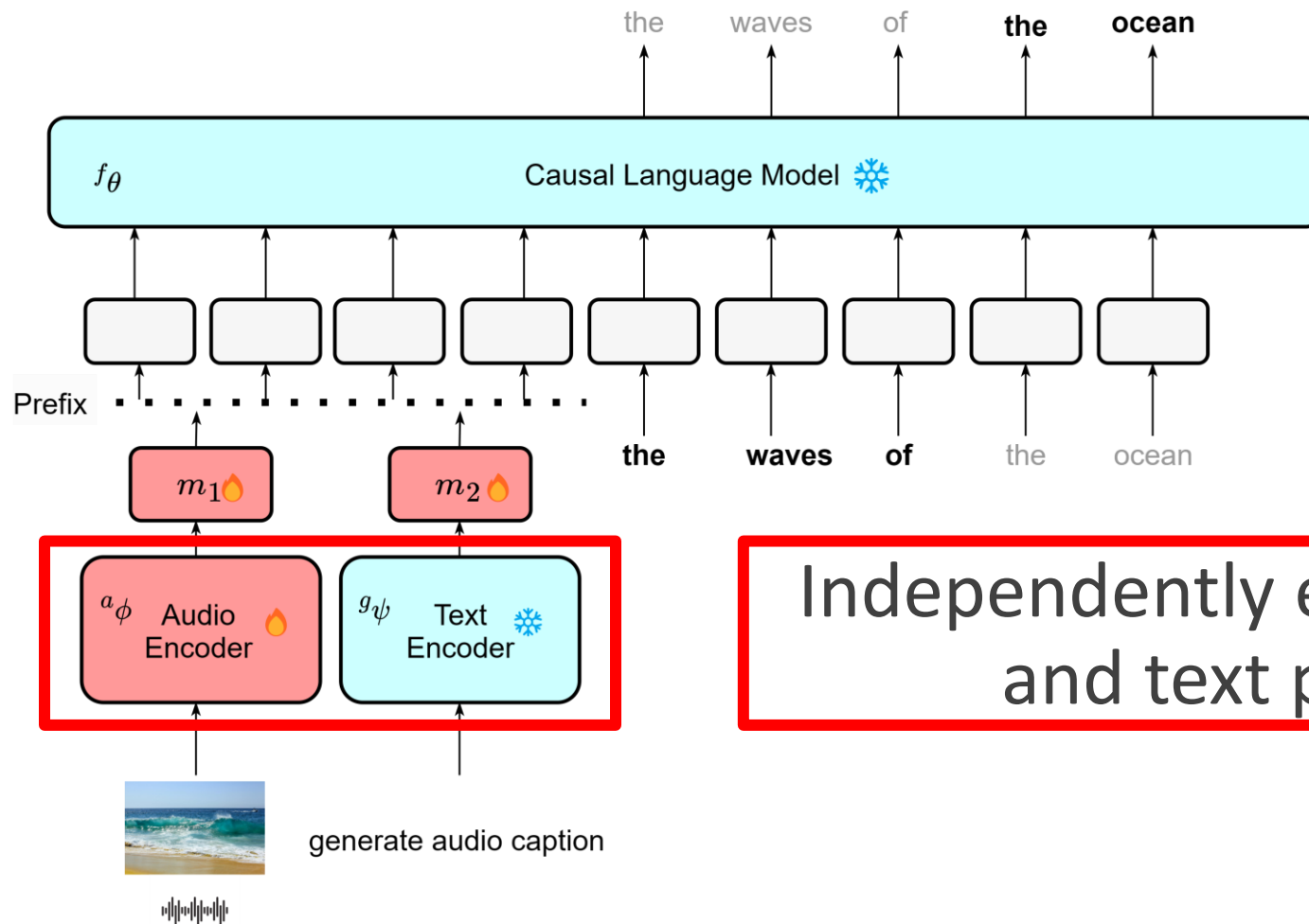
# Model architecture



# Training

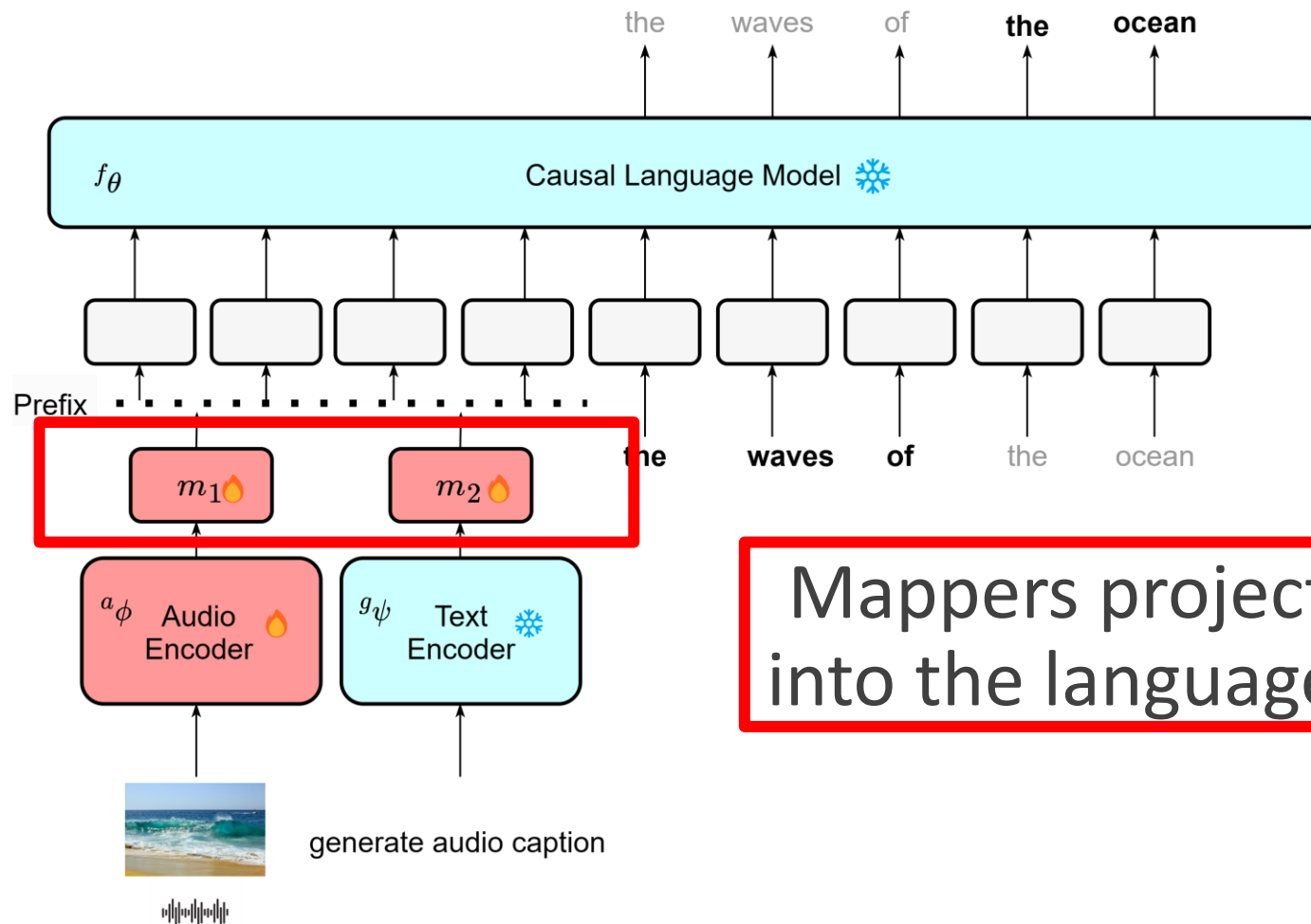


# Training



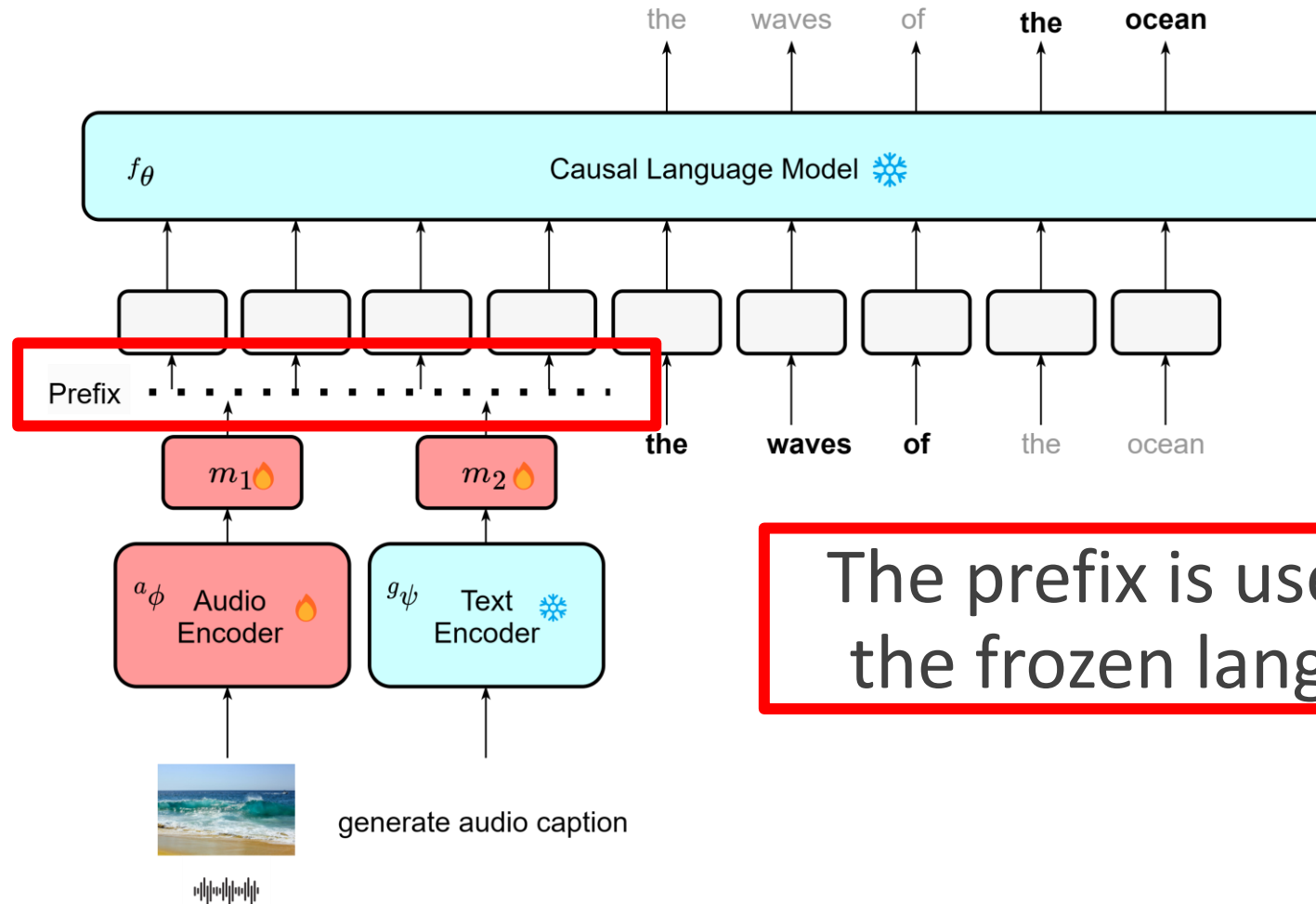
Independently encode audio  
and text prompt

# Training



Mappers project embeddings into the language model space

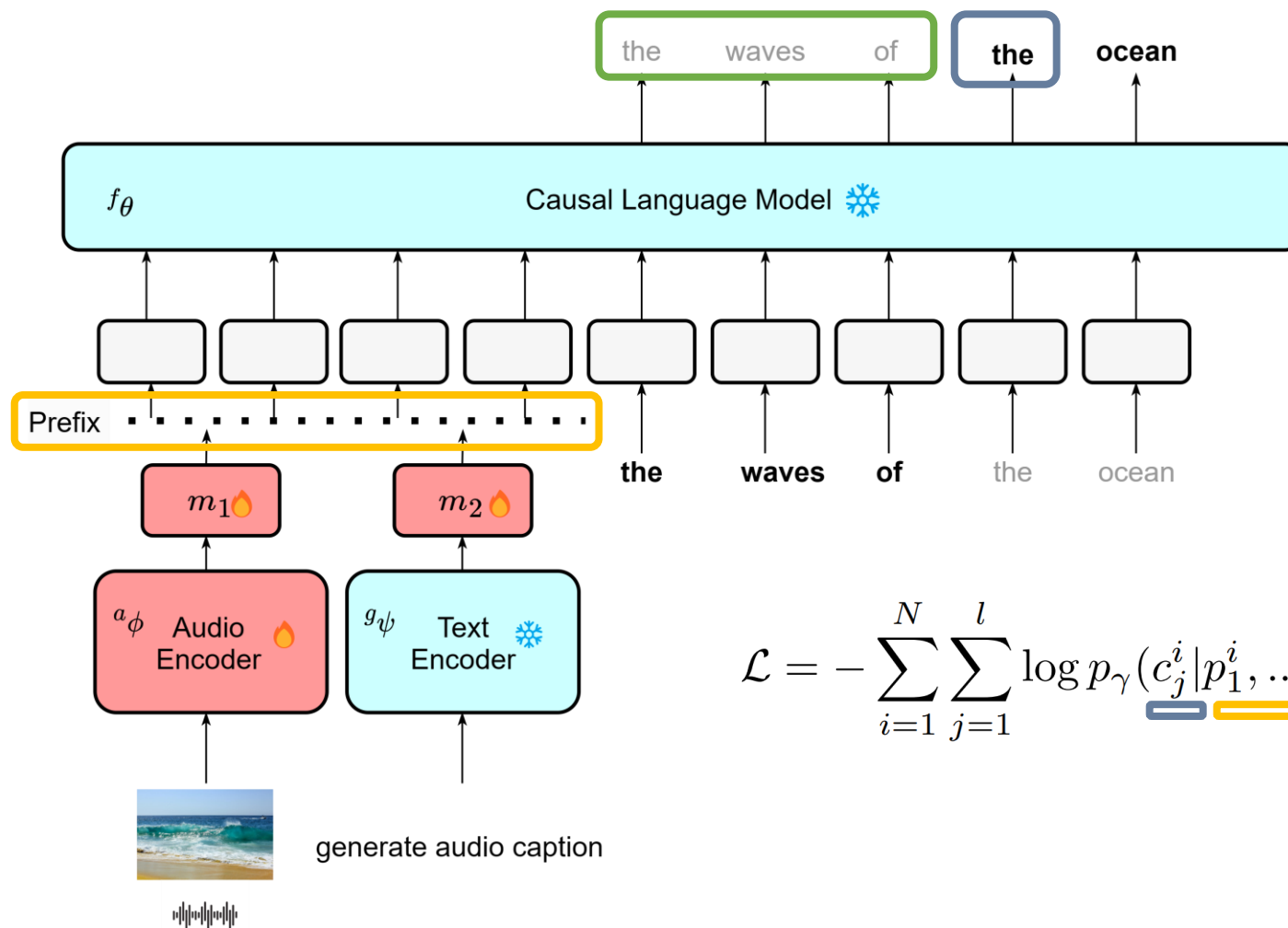
# Training



The prefix is used to prompt the frozen language model

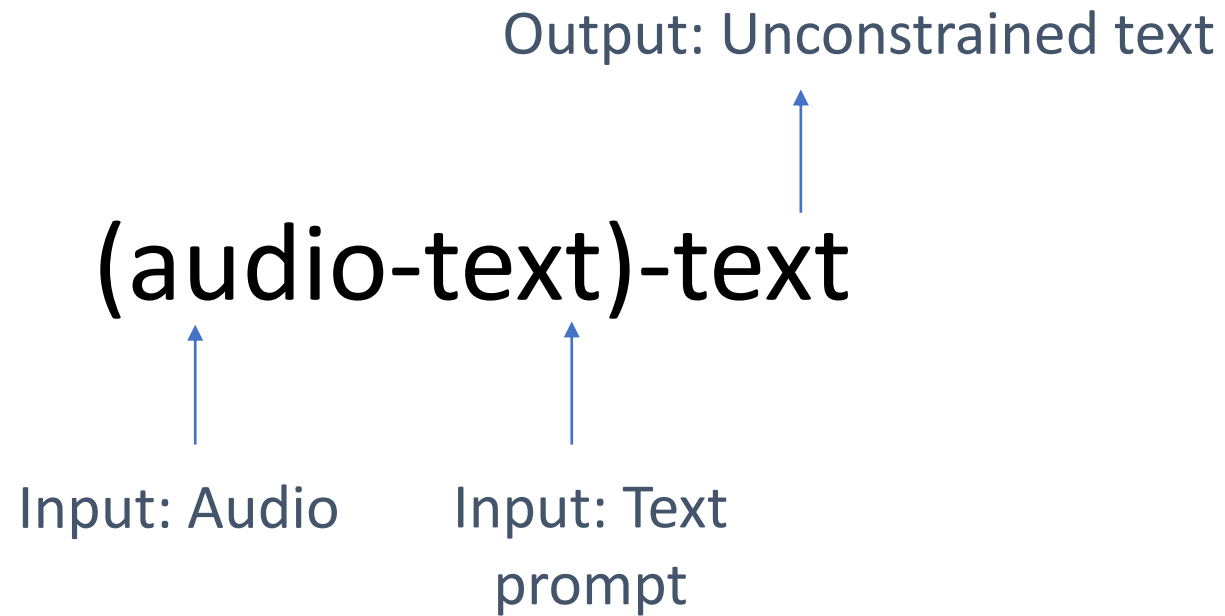


# Training



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

# Audio-task templates for training



**3.4M triplets**

# Benchmarking Pengi on downstream tasks

	Audio Captioning ↑		Audio Q&A ↑	Sound Event Classification ↑			
Model	AudioCaps	Clotho	ClothoAQA	ESC50	FSD50K	US8K	DCASE17 Task 4
CLAP*	<b>X</b>	<b>X</b>	<b>X</b>	0.8916	0.3398	<b>0.7661</b>	<b>0.3387</b>
Pengi	<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
Pengi	<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. ↑	Surveillance ↑
Model	CRE MA-D	RAV DESS	Vocal Sound	ESC50 Actions	SESA
CLAP*	0.1512	0.1692	0.5522	0.508	<b>0.7094</b>
Pengi	<b>0.1846</b>	<b>0.2032</b>	<b>0.6035</b>	<b>0.5277</b>	0.5402

# Benchmarking Pengi on downstream tasks

		Audio Captioning ↑		Audio Q&A ↑	Sound Event Classification ↑			
Model		AudioCaps	Clotho	ClothoAQA	ESC50	FSD50K	US8K	DCASE17 Task 4
CLAP*		<b>X</b>	<b>X</b>	<b>X</b>	0.8916	0.3398	<b>0.7661</b>	<b>0.3387</b>
Pengi		<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
Pengi	<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. ↑	Surveillance ↑
Model	CRE MA-D	RAV DESS	Vocal Sound	ESC50 Actions	SESA
CLAP*	0.1512	0.1692	0.5522	0.508	<b>0.7094</b>
Pengi	<b>0.1846</b>	<b>0.2032</b>	<b>0.6035</b>	<b>0.5277</b>	0.5402

# Benchmarking Pengi on downstream tasks

	Audio Captioning ↑		Audio Q&A ↑	Sound Event Classification ↑			
Model	AudioCaps	Clotho	ClothoAQA	ESC50	FSD50K	US8K	DCASE17 Task 4
CLAP*	<b>X</b>	<b>X</b>	<b>X</b>	0.8916	0.3398	<b>0.7661</b>	<b>0.3387</b>
Pengi	<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
Pengi	<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. ↑	Surveillance ↑
Model	CRE MA-D	RAV DESS	Vocal Sound	ESC50 Actions	SESA
CLAP*	0.1512	0.1692	0.5522	0.508	<b>0.7094</b>
Pengi	<b>0.1846</b>	<b>0.2032</b>	<b>0.6035</b>	<b>0.5277</b>	0.5402

# Contrastive and next-token pretraining

- CLAP can be used for zero-shot close-ended tasks, such as classification and retrieval
- Pengi an Audio-Language model that can perform both open-ended and close-ended downstream tasks
- Can be combined to get better models

# Contrastive and next-token pretraining

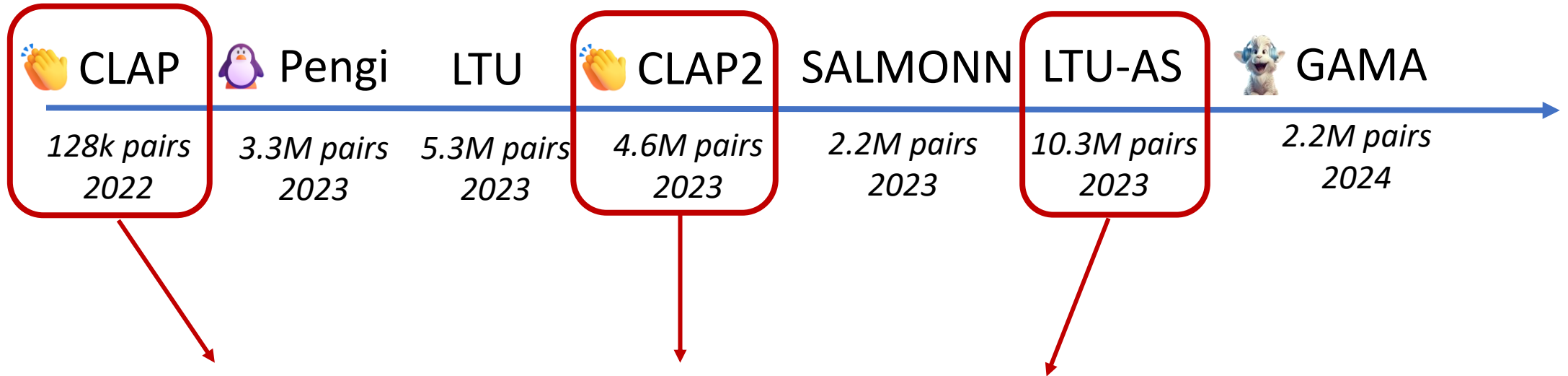
- CLAP can be used for zero-shot close-ended tasks, such as classification and retrieval
- Pengi an Audio-Language model that can perform both open-ended and close-ended downstream tasks
- Can be combined to get better models

# Talk outline

1. Contrastive-Language Audio Pretraining
2. Audio-conditioned next-token prediction
- 3. Deductive Reasoning**

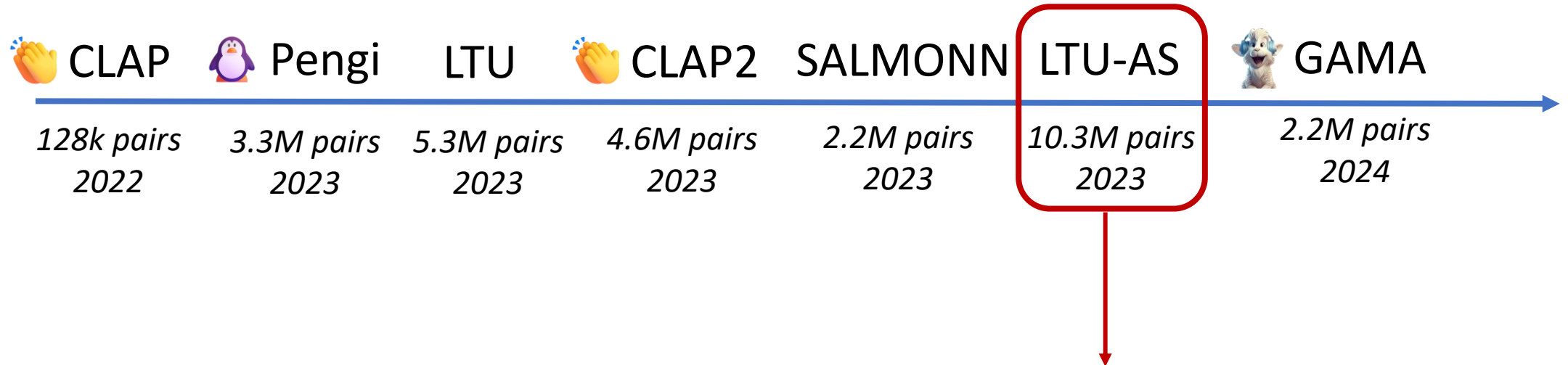


# Scaling data and compute



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



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.



I recommend having a mechanic check your fuel injectors and spark plugs. If left unchecked, this issue could damage the engine.

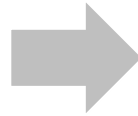


# The audio assistant shows primitive reasoning abilities in the conversation



## 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

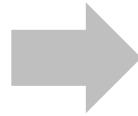


# The audio assistant shows primitive reasoning abilities in the conversation



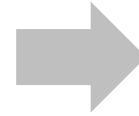
## Perception and understanding

engine knocking sound



## World knowledge

car sounds, engine issues, misfiring cylinder ..

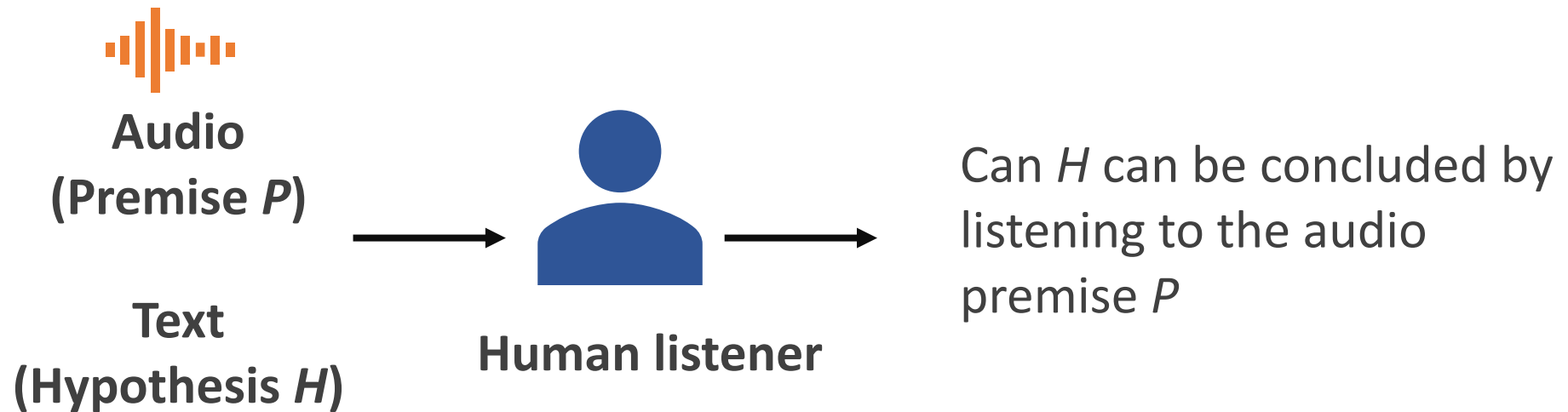


## Logical deduction

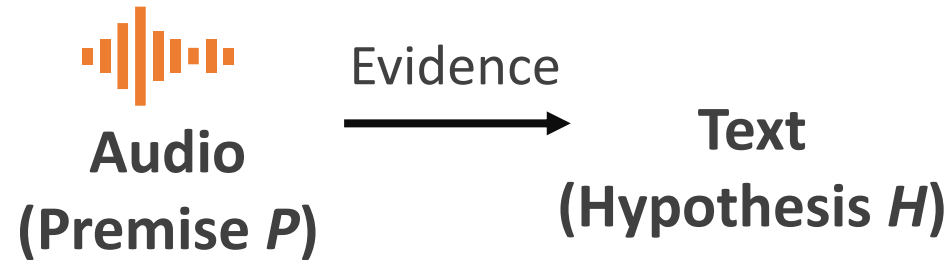
Indicate detonation. It is likely caused by fuel issues or a misfiring cylinder

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

# To evaluate auditory deductive reasoning for humans



# Three possible scenarios

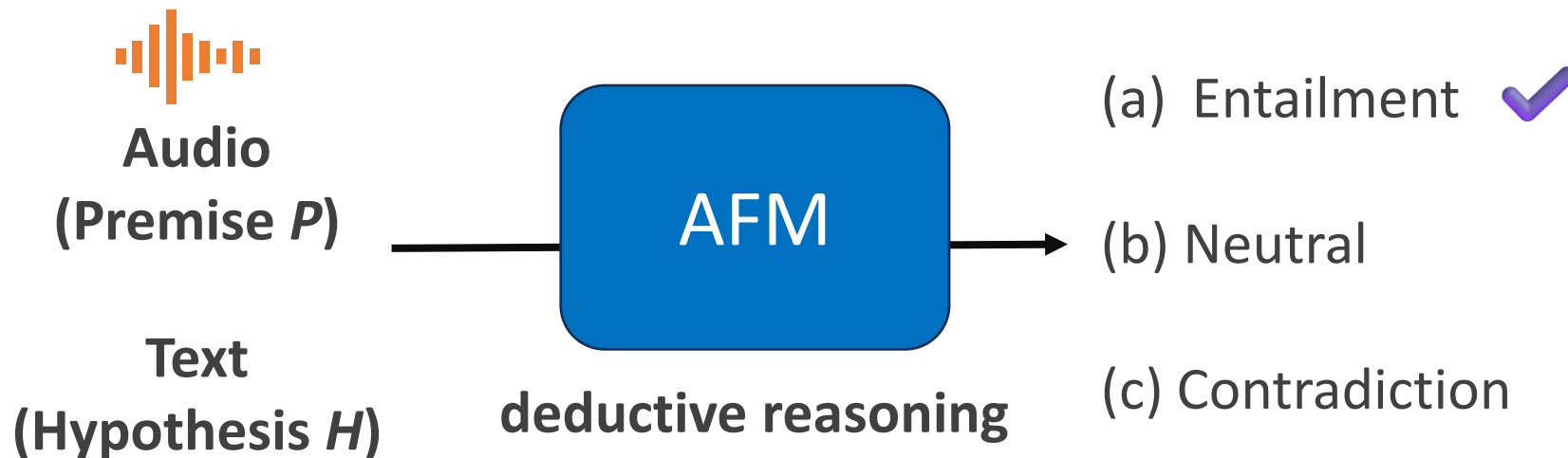


1.  $P$  contains sufficient evidence to affirm the truth of  $H$

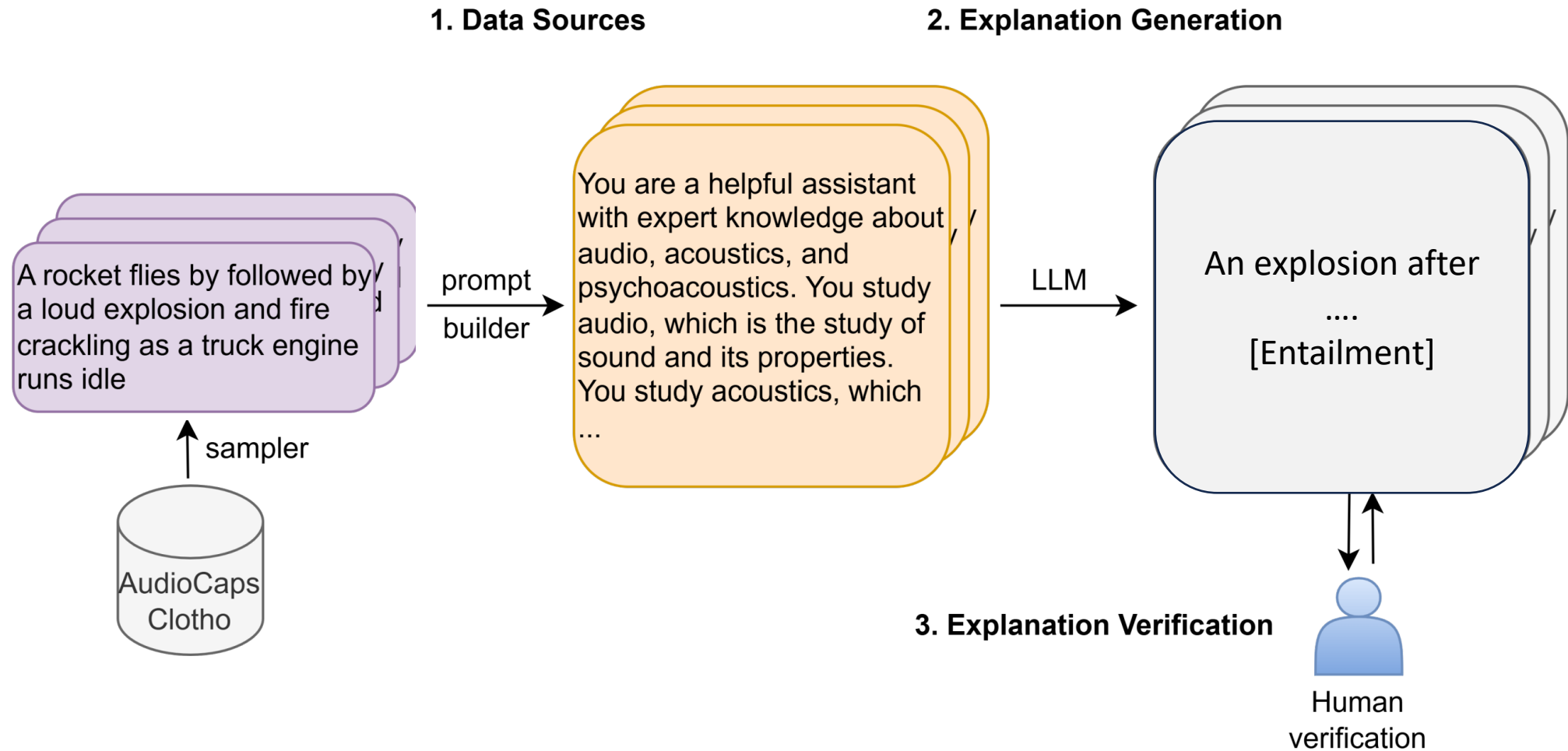
2.  $P$  does not have enough information to either confirm or deny  $H$

3.  $P$  have substantial evidence to deduce that  $H$  is false.

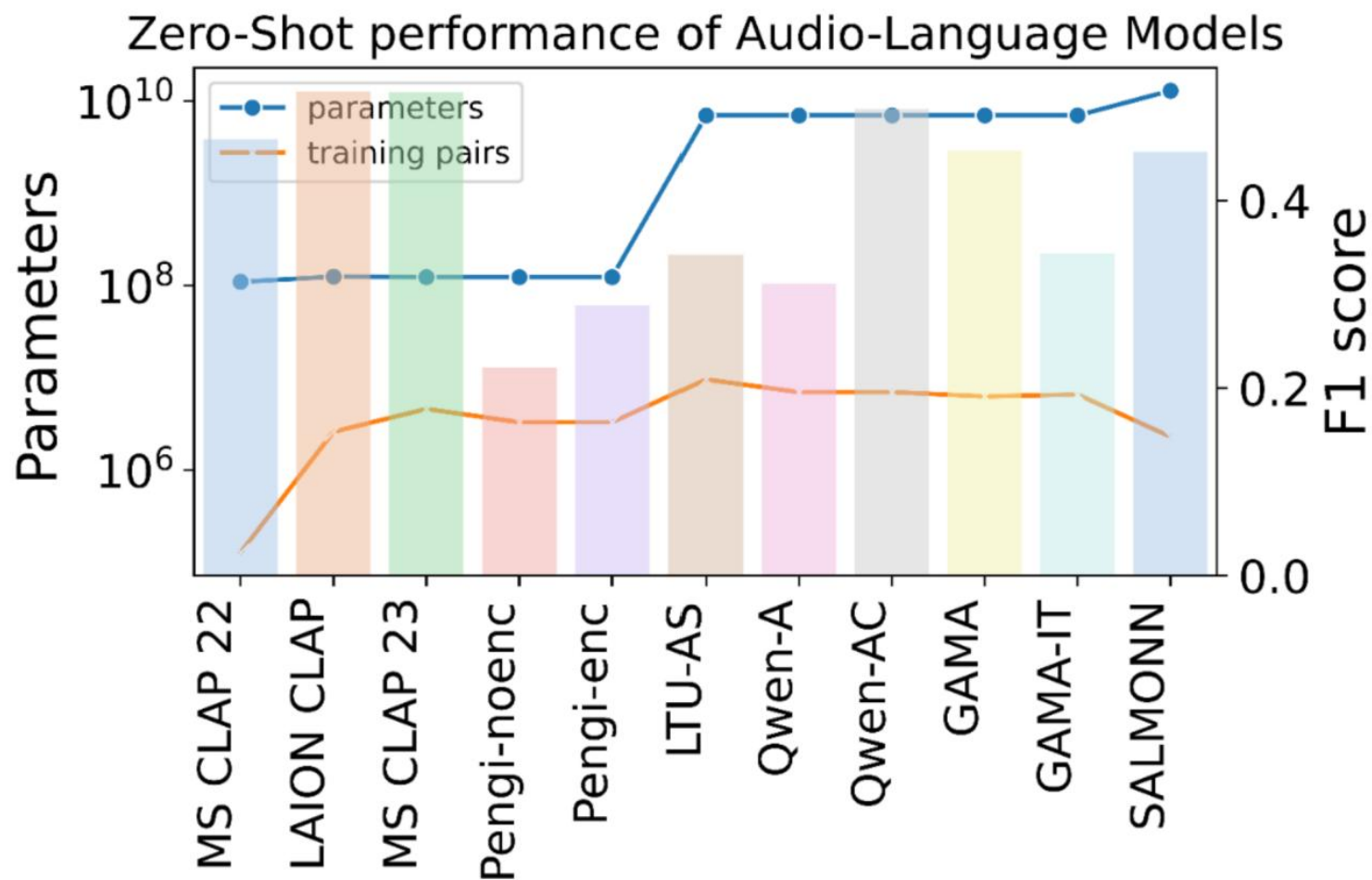
# Introduce Audio Entailment task



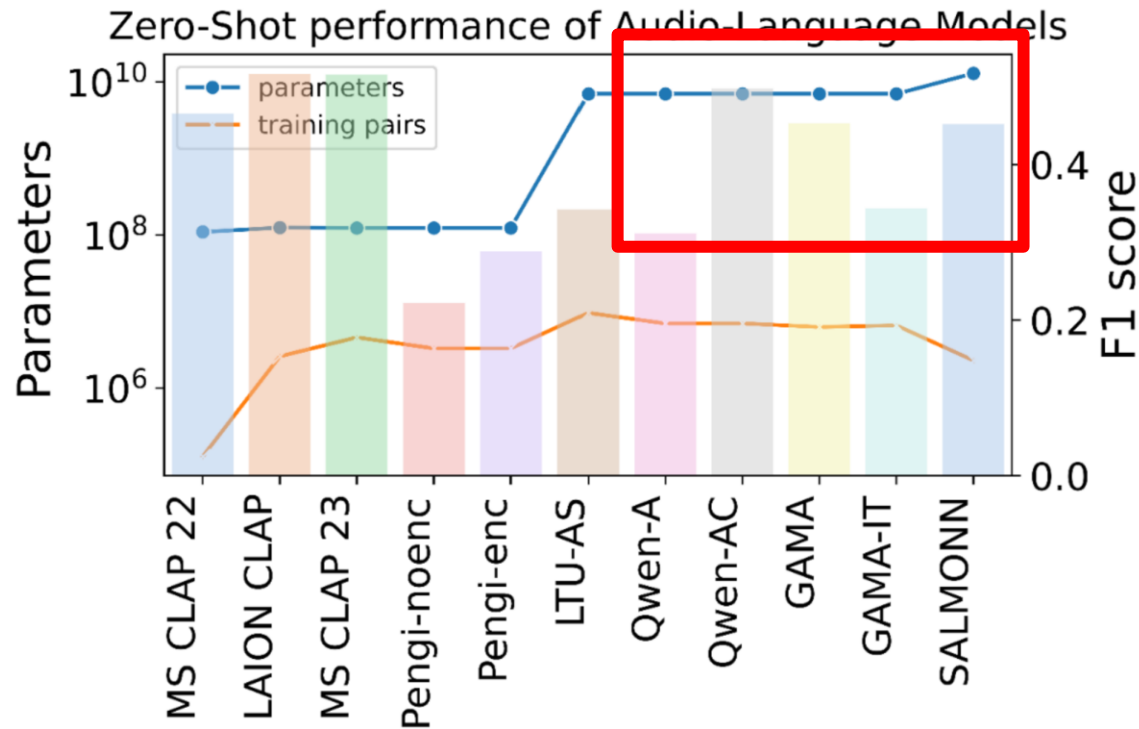
# Audio Entailment task dataset



# Benchmarking AFM on Audio Entailment task



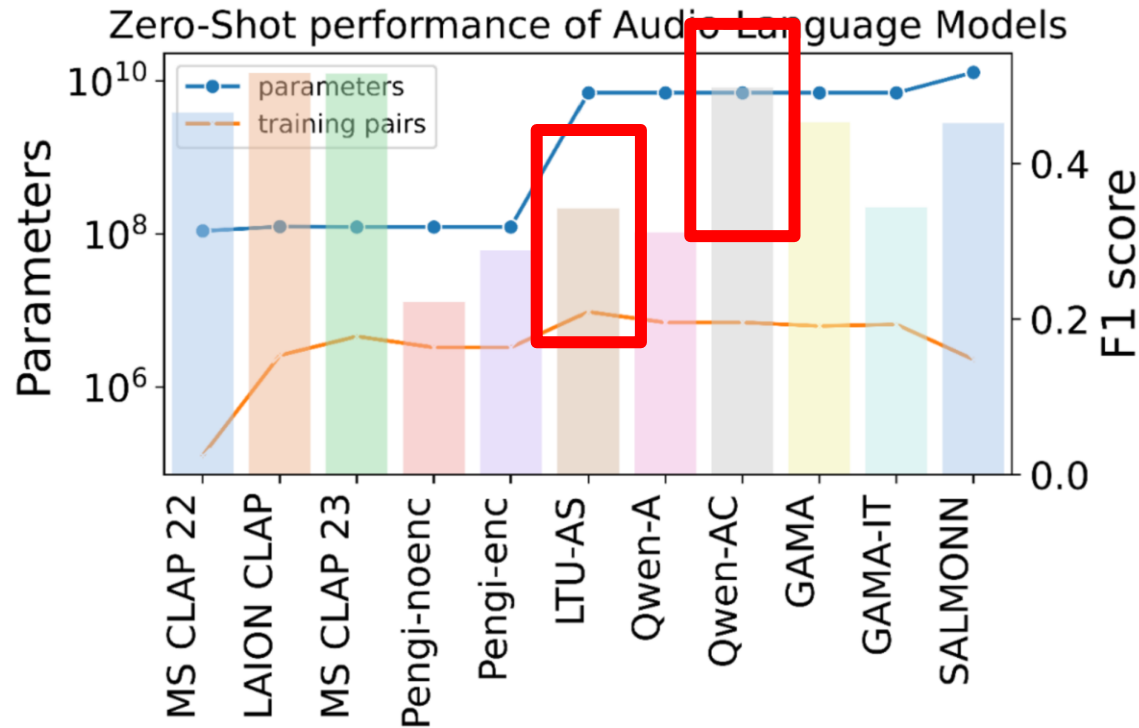
# Findings from benchmarking



(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

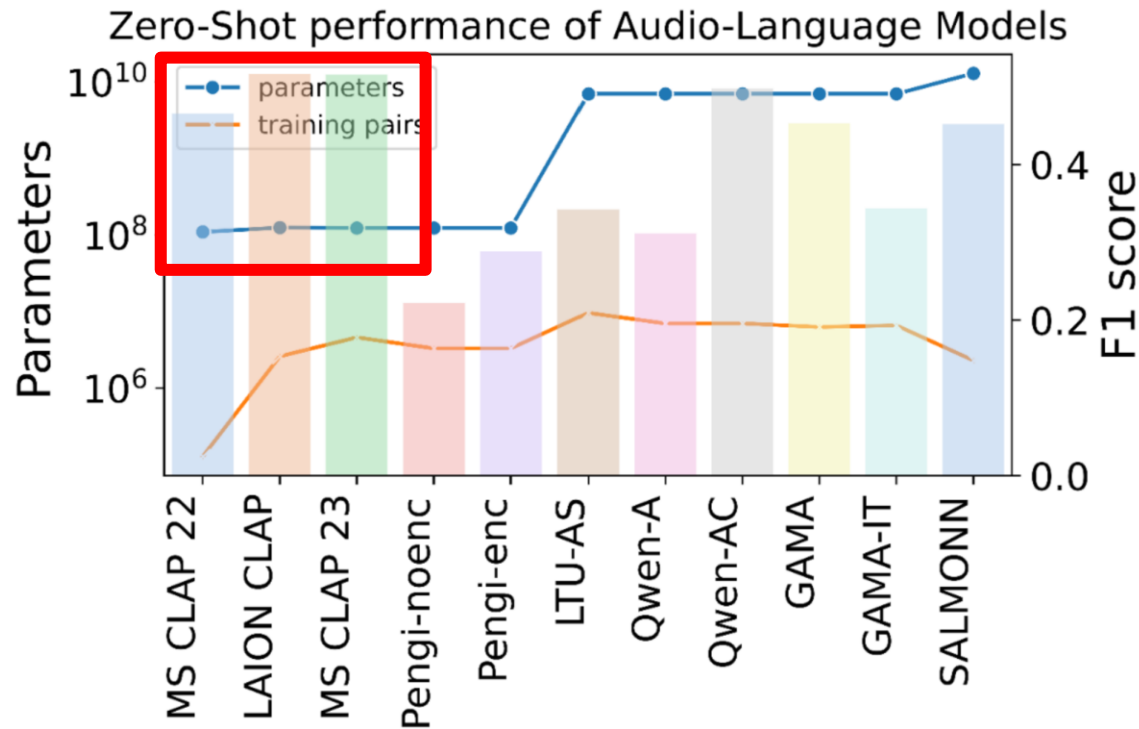


(2) Training AFMs to predict uncertainty improves their ability to detect plausible scenario

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.



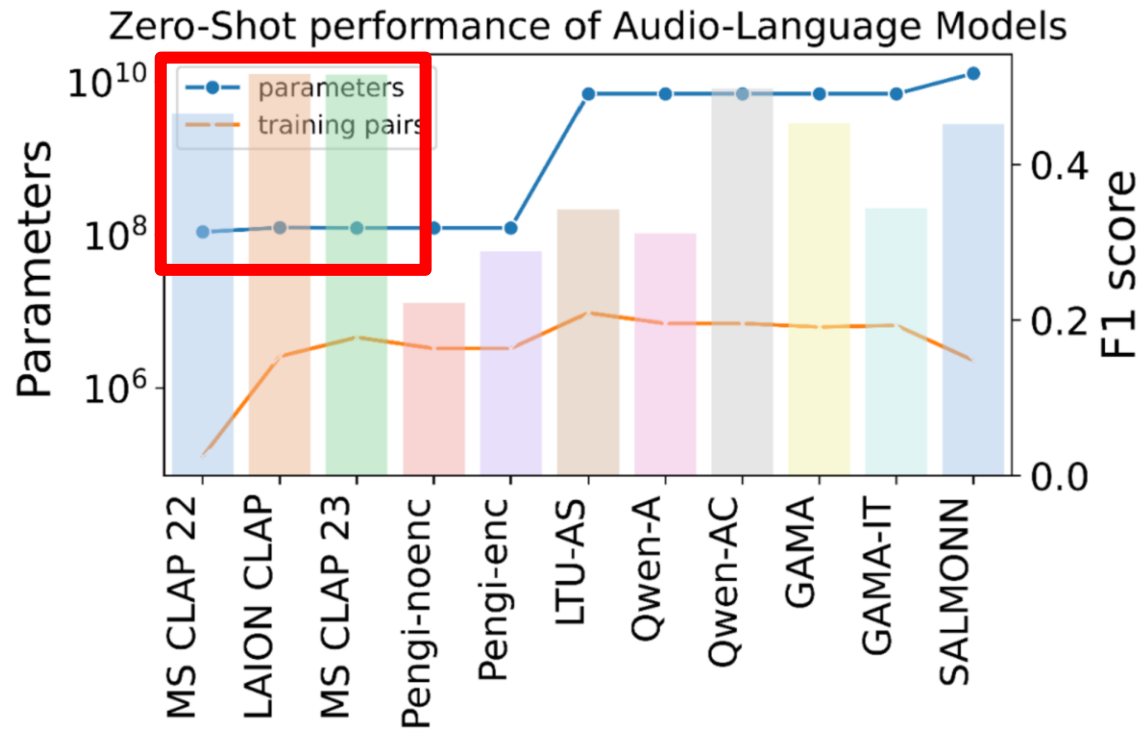
# Findings from benchmarking



(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.

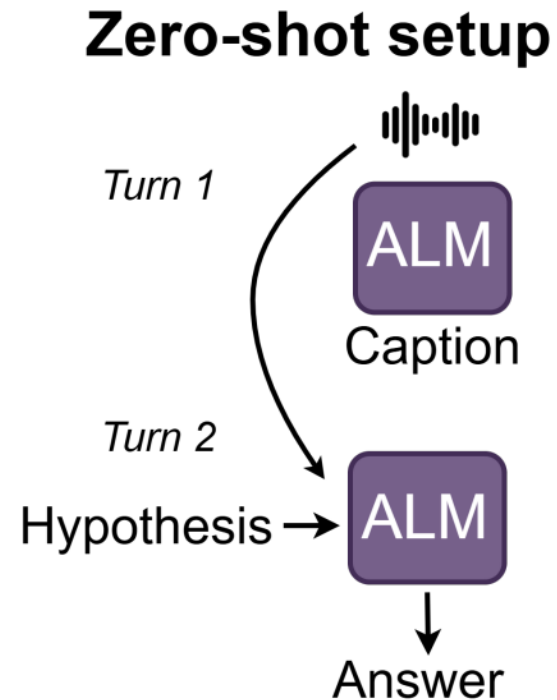
How do we improve deductive reasoning of AFMs at test-time?

# Reduce perception gap

Caption the audio before performing deductive reasoning

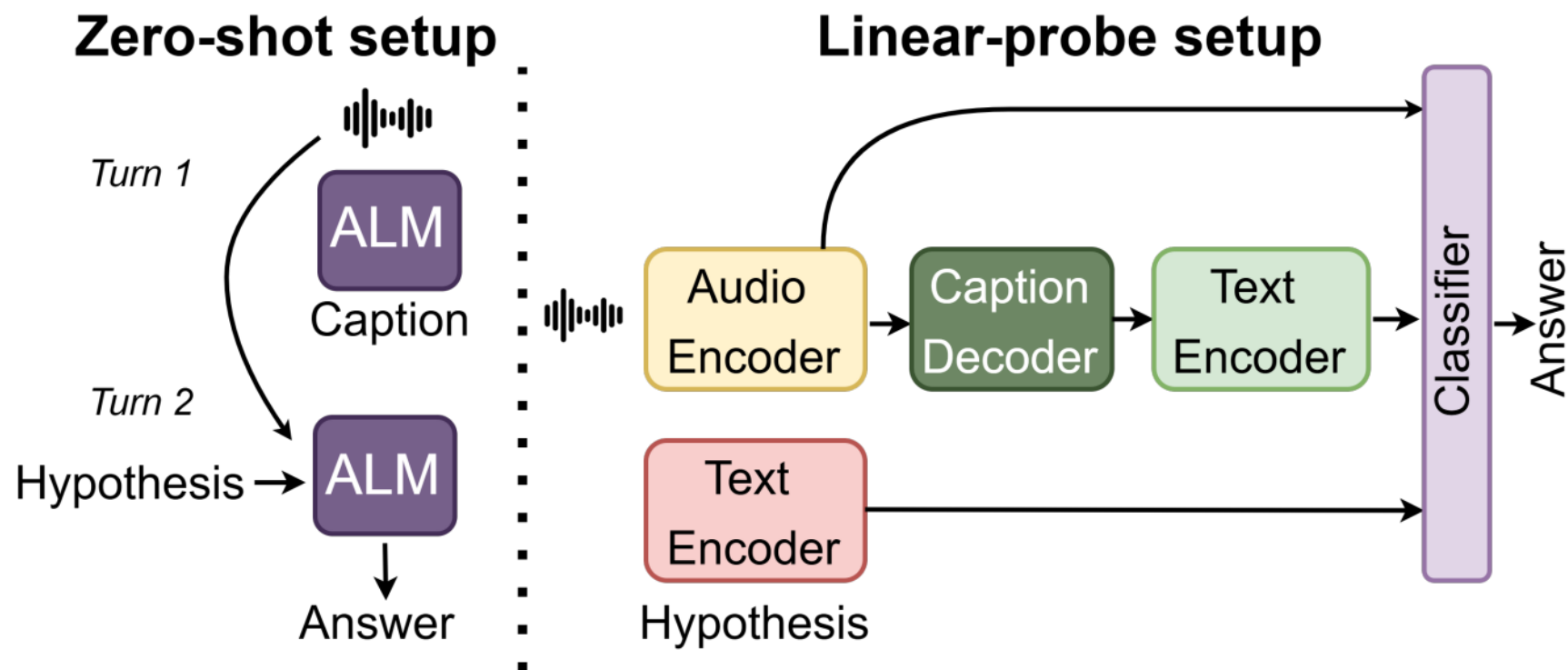
# Reduce perception gap

Caption the audio before performing deductive reasoning



# 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**

# Talk summary + future directions

