

# Advancing Audio Processing in the Age of Large Language Models

Dinesh Manocha

Departments of Computer Science

Department of Electrical & Computer Engineering

University of Maryland, College Park

[dmanocha@umd.edu](mailto:dmanocha@umd.edu)

[https://sakshi113.github.io/audio\\_webpage/](https://sakshi113.github.io/audio_webpage/)

# Collaborators

- Ramani Duraiswamin (UMD)
- FNU Sakshi (UMD)
- **Sreyan Ghosh** (UMD)
- Sonal Kumar (UMD & Adobe)
- Anton Ratnarajah (UMD/Amazon)
- Carl Schissler (UNC/Meta)
- Ashish Seth (UMD)
- Zhenyu Tang (UMD/Meta)
- Industry Collaborators: Adobe, Dolby, NVIDIA

# Audio as a Modality

Humans rely heavily on auditory information in daily life. Auditory cues enhance situational awareness and communication.

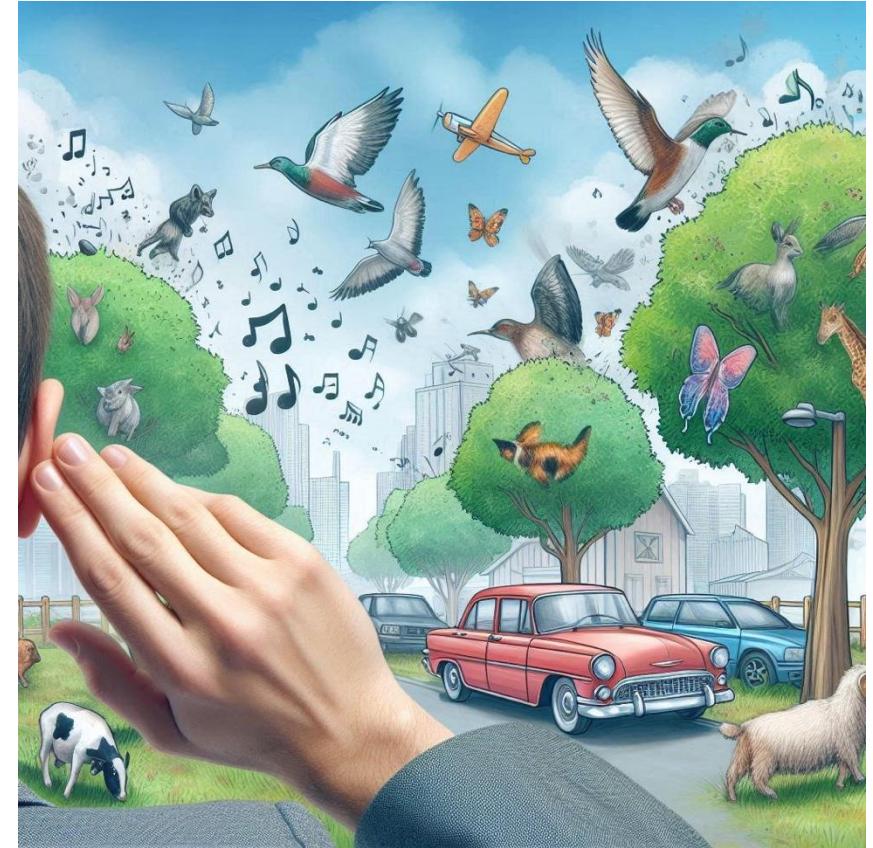
Auditory cues include both verbal and non-verbal audio.



**Verbal audio** includes spoken speech and helps humans communicate.



**Non-verbal audio** includes the non-verbal speech (e.g., sneeze, cough, etc.) and non-speech sounds (e.g., environmental sounds and music). Non-verbal cues are as important as verbal cues

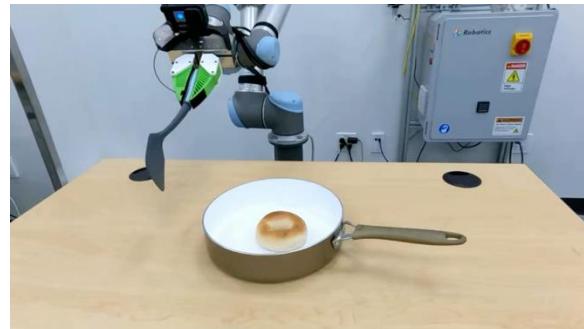


# Audio as a Modality in AI

- AI agents & audio perception** - Audio is essential for creating intelligent, context-aware AI agents. A lot of tasks require agents to understand and reason about the audio cues in the environments.
- Existing tasks can greatly benefit from audio cues** - movie summarization, object detection, etc.
- Audio can improve immersiveness of generated content**  
- enhances immersiveness, creating richer and more engaging experiences that complement visual and textual information.



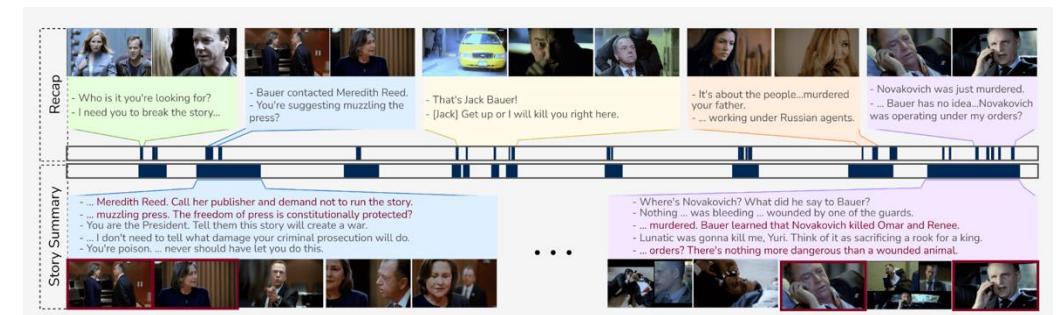
Generated videos can be more immersive with videos!



An example of improving imitation learning in robots without auditory cues. Example from ManiWAV (Liu et al., 2024)



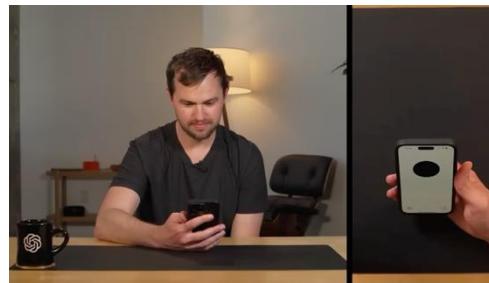
Agents that can hear and speak understand and communicate better with humans.



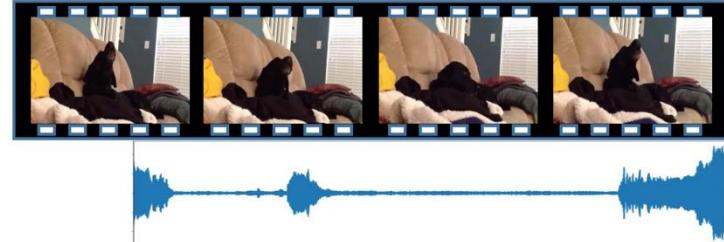
Movie summarization using only videos misses important information hidden in sound effects and human communication!

# Examples of Improving AI systems with Audio Perception

- Scene understanding has improved incredibly with audio integration:
  - Improved ego-centric perception
  - Improved scene captioning
- GPT-4o is integrated with hearing and speaking capabilities.
- Visual-only tasks have shown to benefit with audio integration
  - Audio-Visual Object detection
  - Video Question-Answering with audio cues
  - Robot Navigation with audio-visual cues

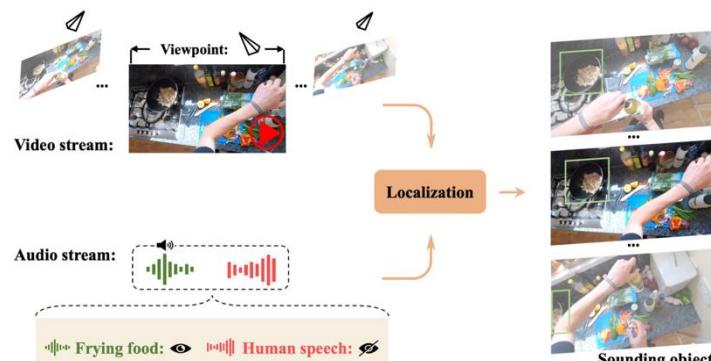


**GPT4o can now hear and speak!** Audio integration has brought a leap in GPT's capabilities.

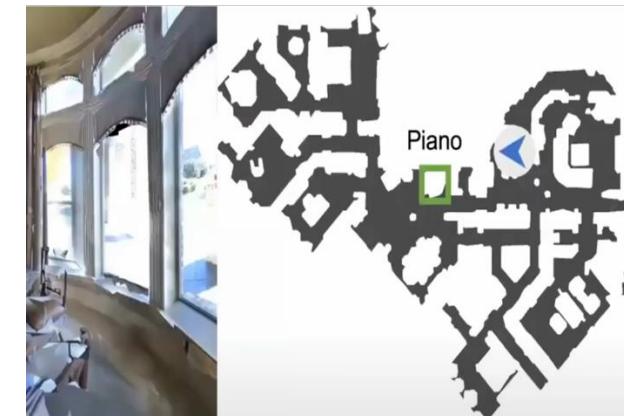


## Audio-Visual Scene Captioning

**Caption w/o Audio:** A black dog lies on the sofa.  
**Caption with Audio:** In the living room, a black dog lies on the sofa barking, with the sound of a police car in the background.



**Egocentric Audio-Visual Object Detection.**  
(Huang et al., 2023)

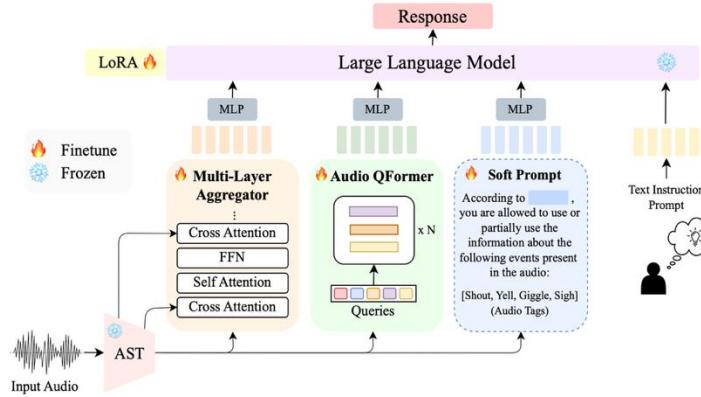


**Robot navigation with audio-visual cues.**  
(Huang et al., 2023)

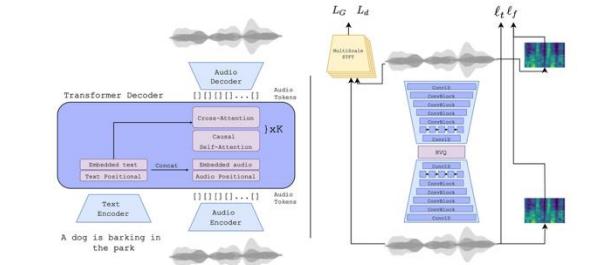
# Recent Advances in Audio Processing

## • Audio Understanding Systems:

- Encoders such as Audio Spectrogram Transformer (Gong et. Al 2021), HTSAT (Chen et al. 2022), etc.
- Audio-Language models such as CLAP (Elizalde et. al 2022), CompA (Ghosh et al. 2023), etc.
- Audio LLMs such as LTU (Gong et. al 2023), GAMA (Ghosh et al. 2024), etc.

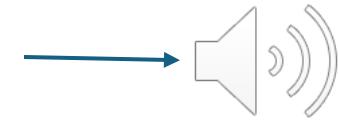


LLMs that can reason about input speech and sounds and answer queries (GAMA, Ghosh et al. 2024 ).



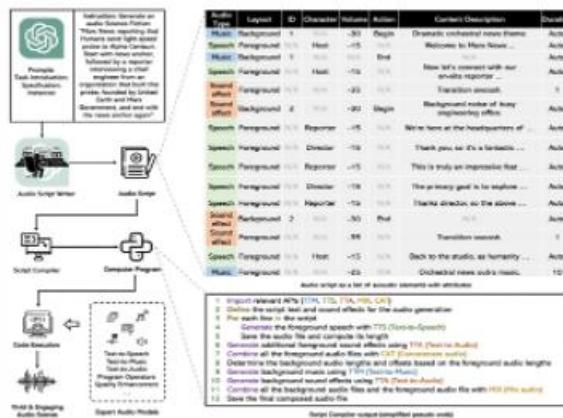
Diffusion models that can generate audio from a query text (AudioGen, Kreuk et al., 2022).

Recreate a gentle rainfall with distant thunder.



## • Audio Generation Systems:

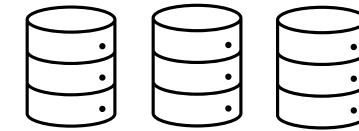
- Diffusion-based Acoustic Event Generation Models such as AudioGen (Kreuk et al. 2022), etc.
- Speech Generation Models such as VALLE (Wang et al. 2023), etc.
- Audio Story Telling Systems such as WavJourney (Liu et al. 2023), etc.



An amalgamation of understanding and generation for audio storytelling (WavJourney, Liu et al., 2023)

# Current Challenges in Audio Processing

- **Limited Audio Datasets:** Audio data is scarce compared to other modalities, with minimal availability for certain tasks. Many datasets are restricted or lack real-world diversity.
  - **Lack of Reasoning-Centric Data:** Most datasets focus on recognition rather than complex reasoning. Unlike language and vision, distilling larger models for reasoning remains a challenge.
  - **Lack of Long Audio Processing Data:** Open-source datasets typically contain **5–10s** clips, with no datasets dedicated to long-form **sounds and music** beyond ASR.
- **Weak Audio Representations:**
  - State-of-the-art **audio encoders achieve ~50% on AudioSet**, far below **95%+ on ImageNet**.
  - Models struggle with **compositional audio structure** (e.g., event order) and **linguistic variations** (e.g., "chopper" vs. "helicopter").
  - Speech, sounds, and music are still treated **separately** rather than holistically.
- **Lack of Good Evaluation Benchmarks:** Most benchmarks remain recognition-focused and are clean and simple, limiting progress in real-world **understanding and reasoning** tasks. 25+ reasoning benchmarks exist for vision and language, but only a few for audio.



Language Datasets  
Largest Dataset - Billions of tokens



Vision Datasets  
Largest Dataset - 5B+ I-T pairs



Speech Datasets  
Largest Dataset - 60k+ Hours



Non-verbal Audio Datasets  
Largest Dataset - ~1k Hours



**Simple Event Detection:** What are the acoustic events in the given audio.

**Complex Instruction:** From the speech and impact sounds, deduce the size and characteristics of the room in which these events occur.

# How are the state-of-the-art audio models doing?

- **Compositional Understanding for Audio Processing:** Real-world audio is often **compositionally complex**. Understanding and generative models trained on clean audios find it difficult to generalize to real-world audios.
  - More compositional audio data
  - Better model architectures
  - Better training algorithms
  - Better audio representations for generative models
- **Advanced Reasoning in Audio:** Current state of the art models **struggle to answer complex questions** on audio which require advanced reasoning and answering using world knowledge.
  - Advanced evaluation benchmarks
  - Better reasoning-centric data
- **Long and Multi Audio Processing:** Most audio models can perceive a **single audio** with a max of **30 secs** in length (with 10 secs being the average).
  - Long audio encoders
  - Long audio representation learning
  - Long audio datasets and benchmarks
  - Multi-audio reasoning strategies



**Caption:** A dog barking followed by a person speaking – generated by Stable Audio



**Question:** Based on various sounds, infer what the man might be announcing.

**Answer by LTU (Gong et. al):** It is not possible to infer.

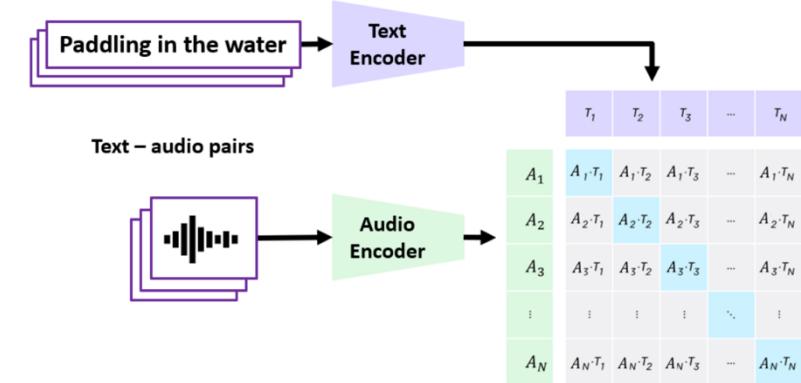


**Question:** What indicates a sudden moment of fear in the audio?

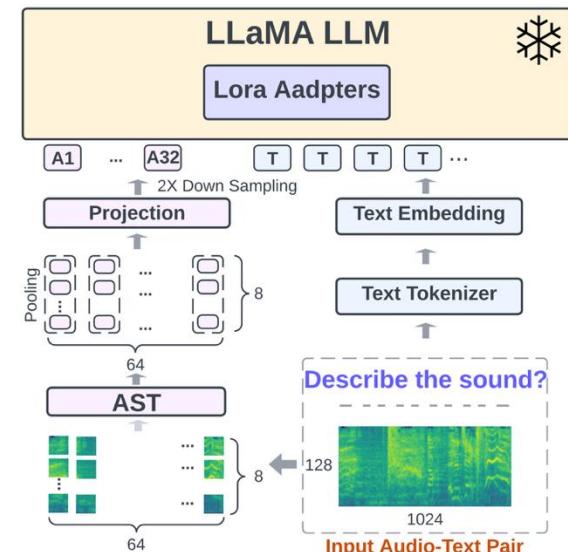
**Answer by LTU:** There is no such moment in the audio.

# (Large) Audio Language Models (LALMs)

- **What are Audio-Language Models (ALMs)?**
  - **Goal:** understand, generate, and interact with audio data
  - Audio processing is challenging due to data scarcity, but **language is abundant.**
  - ALMs learn a **shared latent space between audio-language** modalities.
- **Pre-LLM Era: CLAP & Contrastive Learning**
  - CLAP (inspired by CLIP) was a key pre-LLM ALM, using contrastive learning to align audio and text.
  - Useful for **classification, audio-to-text retrieval, and text-to-audio retrieval.**
- **Post-LLM Era: LALMs**
  - **LALMs integrate audio encoders with LLMs**, unlocking deeper **audio perception and reasoning**.
  - Examples: **GAMA, LTU, Qwen, AudioFlamingo, SALMONN**.
  - Capable of **captioning, reasoning and open-ended QA**.

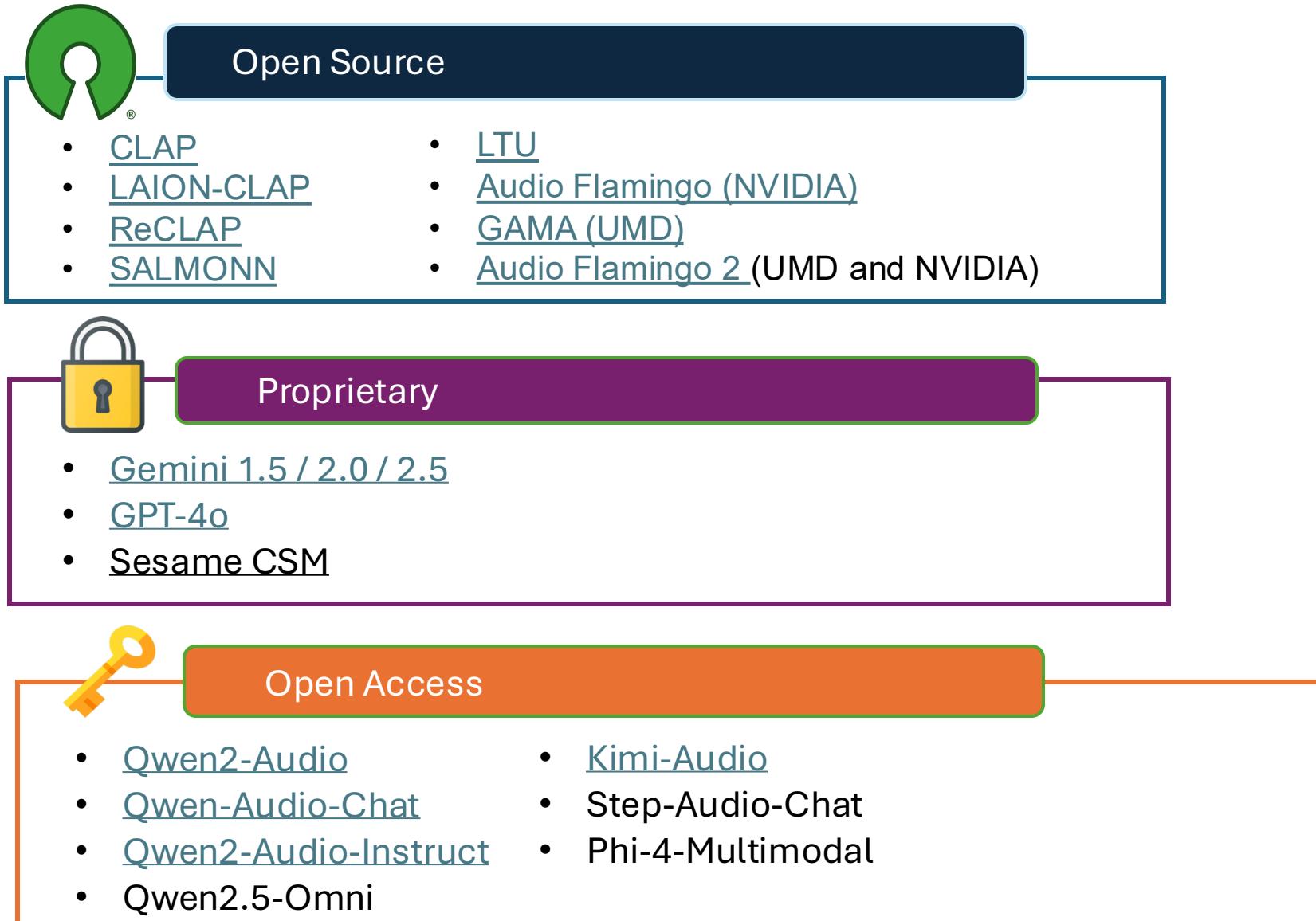


CLAP model learned with contrastive audio-text pairs.



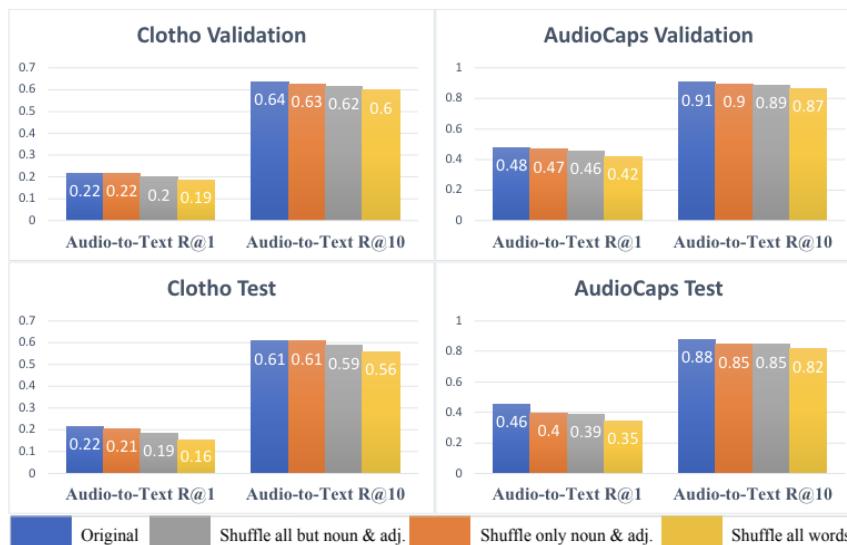
LALMs with audio encoders integrated with LLMs.

# Current State-of-the-Art Audio (Large) Language Models

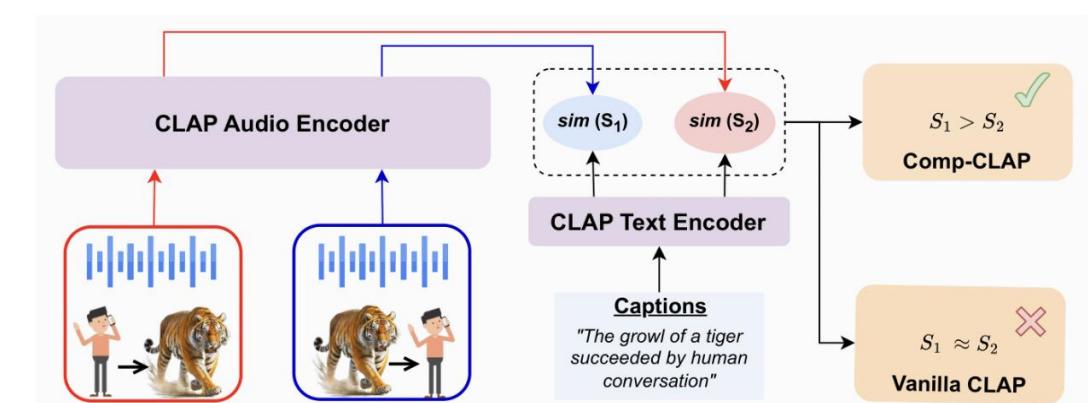


# CompA: Addressing the Gap in Compositional Reasoning in Audio-Language Models (ICLR 2024)

- **Compositional Reasoning:** Understanding the relationship between text in captions and the corresponding content of the audio is a fundamental goal of audio processing. Different word orders should correspond to differently perceived audio.
- Current benchmarks are insufficient for evaluating compositional reasoning.
- Current CLAP-like Audio-Language models lack compositional reasoning.



Model performance does not drop on current retrieval benchmarks with words shuffled.



Current ALMs are not good at compositional reasoning.

# CompA: CompA Benchmark

We propose CompA, the first suite of benchmarks for evaluating compositional reasoning in ALMs. **CompA-order** evaluates an ALM's capability to understand the order of occurrence between multiple acoustic events in an audio. **CompA-attribute** evaluates an ALM's capability to understand attribute-binding for multiple acoustic events in an audio.

**CompA-Order**

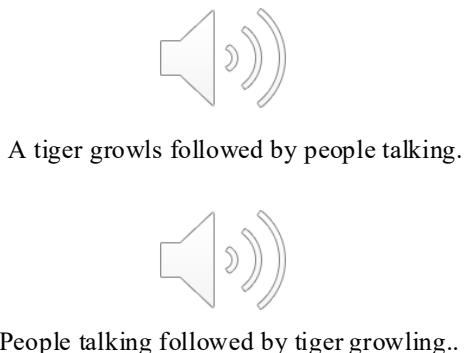
→ Order   \* Overlap

**Captions**

	The growl of a tiger succeeded by human conversation	Human conversation succeeded by the growl of a tiger	The growl of a tiger amidst human conversation
Order	✓	✗	✗
Overlap	✗	✓	✗
	✗	✗	✓

**Audio**

The interface shows three pairs of audio waveforms. The first pair shows a tiger growl followed by a human conversation. The second pair shows a human conversation followed by a tiger growl. The third pair shows a tiger growl overlapping with a human conversation.



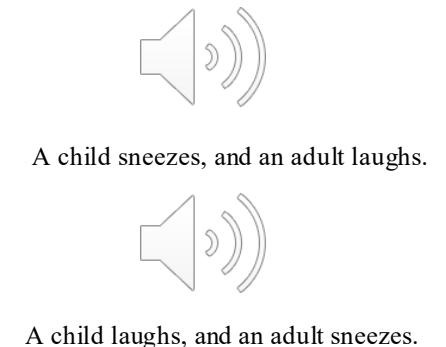
**CompA-Attribute**

**Captions**

	A baby cries while a woman laughs	A woman cries while a baby laughs
Cries	✓	✗
Laughs	✗	✓

**Audio**

The interface shows two pairs of audio waveforms. The first pair shows a baby crying while a woman laughs. The second pair shows a woman crying while a baby laughs.



# CompA: CompA Benchmark Evaluation

$$f(C_0, A_0, C_1, A_1) = \begin{cases} 1 & \text{if } s(C_0, A_0) > s(C_1, A_0) \\ & \text{and } s(C_1, A_1) > s(C_0, A_1) \\ 0 & \text{otherwise} \end{cases}$$

Text score.

$$g(C_0, A_0, C_1, A_1) = \begin{cases} 1 & \text{if } s(C_0, A_0) > s(C_0, A_1) \\ & \text{and } s(C_1, A_1) > s(C_1, A_0) \\ 0 & \text{otherwise} \end{cases}$$

Audio score.

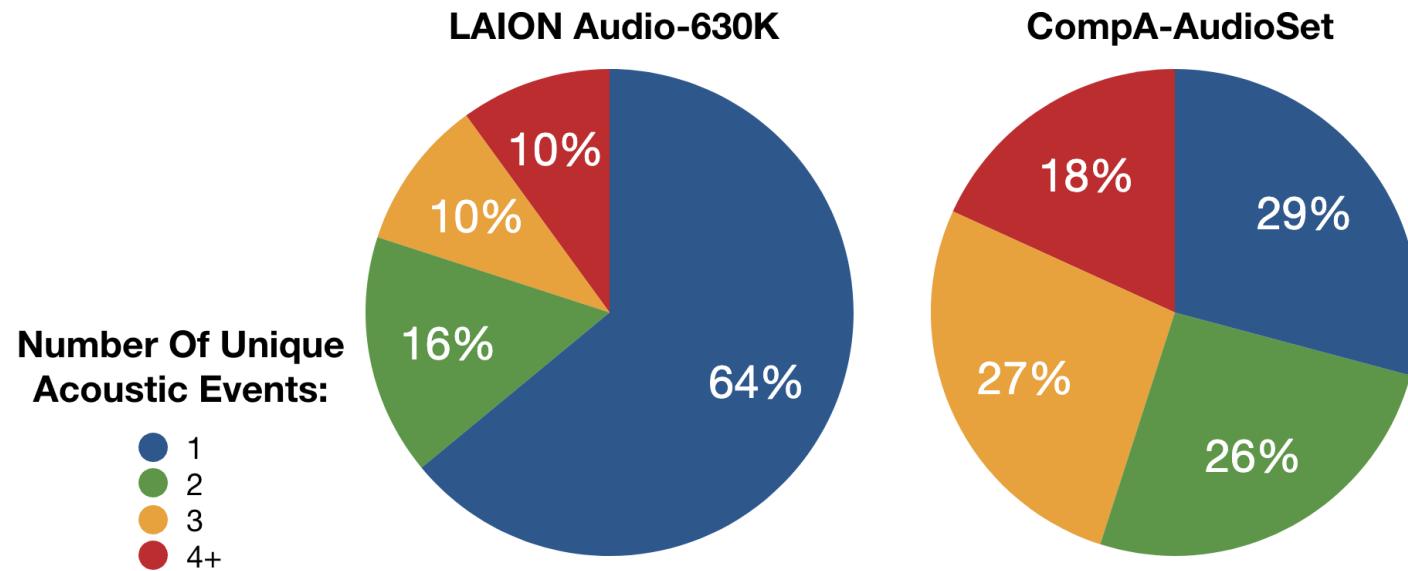
$$h(C_0, A_0, C_1, A_1) = \begin{cases} 1 & \text{if } f(C_0, A_0, C_1, A_1) \\ & \text{and } g(C_0, A_0, C_1, A_1) \\ 0 & \text{otherwise} \end{cases}$$

Combined score.

Model	Text	CompA-order		CompA-attribute		
		Audio	Group	Text	Audio	Group
Human	90.60	91.20	87.40	80.30	82.40	79.80
Random	19.70	19.70	16.67	25.0	25.0	16.67
MMT	$19.90 \pm 1.30$	$6.85 \pm 1.90$	$3.90 \pm 1.95$	$29.59 \pm 1.03$	$4.69 \pm 2.29$	$3.12 \pm 1.76$
ML-ACT	$21.89 \pm 1.75$	$8.00 \pm 0.80$	$4.35 \pm 1.25$	$31.63 \pm 1.46$	$5.11 \pm 2.02$	$3.75 \pm 0.86$
CLAP	$22.80 \pm 2.15$	$8.35 \pm 1.40$	$4.70 \pm 2.20$	$33.27 \pm 0.72$	$6.14 \pm 1.37$	$4.66 \pm 2.08$
CLAP-LAION	$24.0 \pm 1.10$	$9.25 \pm 1.15$	$5.50 \pm 0.80$	$34.78 \pm 1.45$	$6.52 \pm 1.47$	$5.07 \pm 1.62$

Current models perform poorly.

# CompA: CompA-661k

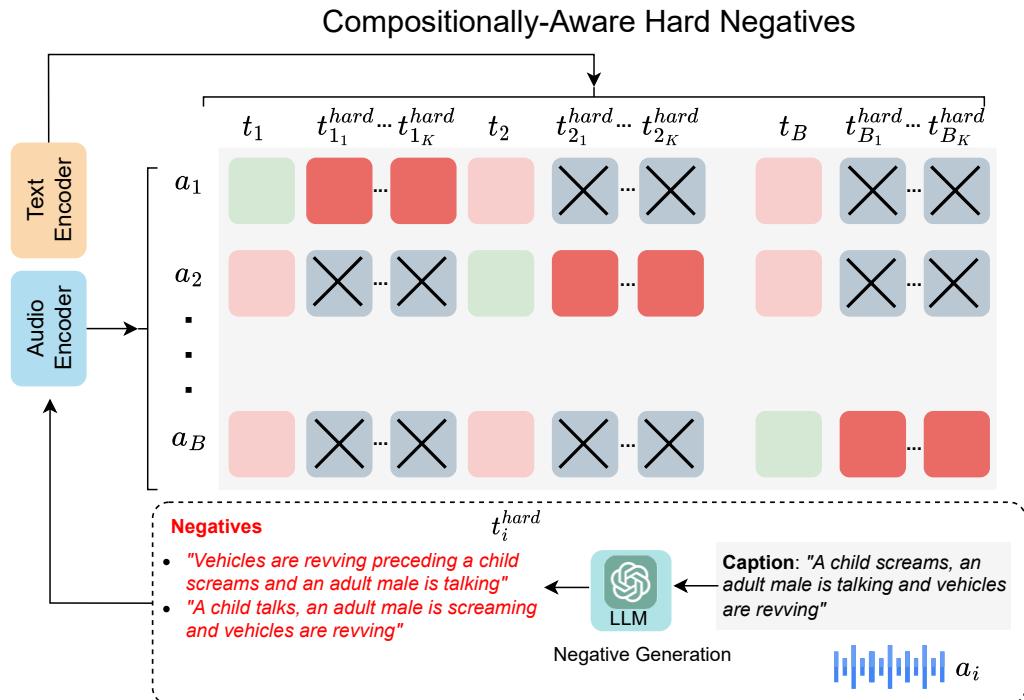


Overview of **CompA-661k**, combining open-source datasets with our proposed **CompA-AudioSet**.  
CompA-AudioSet improves compositional balance compared to **LAION Audio-630k**.

Model	CompA-order			CompA-attribute		
	Text	Audio	Group	Text	Audio	Group
Human	90.60	91.20	87.40	80.30	82.40	79.80
Random	19.70	19.70	16.67	25.0	25.0	16.67
MMT	19.90	6.85	3.90	29.59	4.69	3.12
ML-ACT	21.85	8.00	4.35	31.63	5.11	3.75
CLAP	22.80	8.35	4.70	33.27	6.14	4.66
CLAP-LAION	24.0	9.25	5.50	34.78	6.52	5.07
CLAP ( <i>ours</i> )	33.75	15.75	11.50	42.40	20.50	14.75

# CompA: CompA-CLAP

**Compositionally-Aware Hard Negative Training:** We introduce two new algorithms to enhance compositional representation learning in CLAP. First, we modify the CLAP objective to incorporate compositionally-aware hard negatives. In this approach, each audio sample in the batch is assigned additional hard negatives that are not shared across other batches, improving the model's ability to distinguish fine-grained compositional variations.

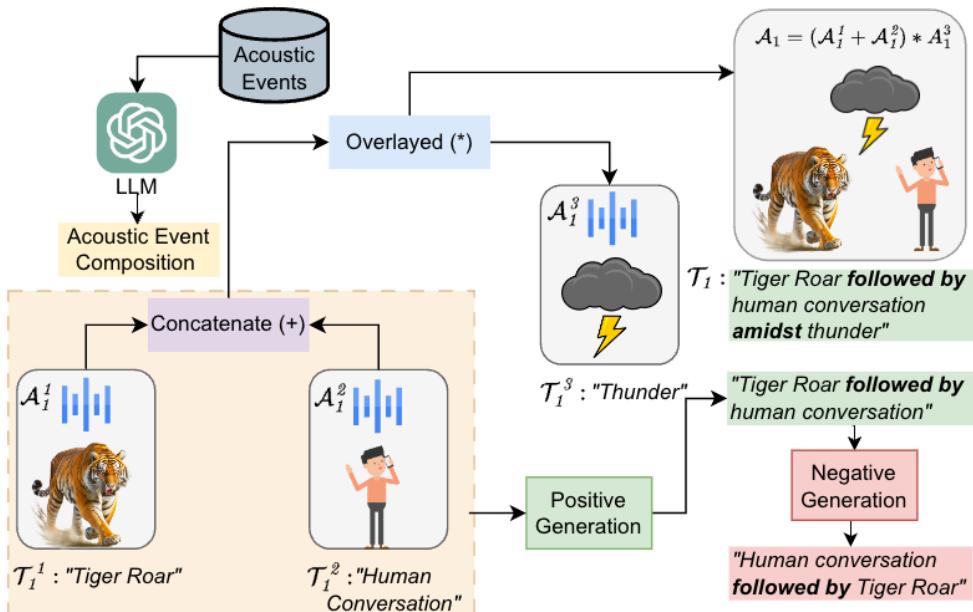


$$\ell_i^{t-2-a} = -t_i^\top a_i / \sigma + \log \sum_{j=1}^B \exp(t_i^\top a_j / \sigma)$$

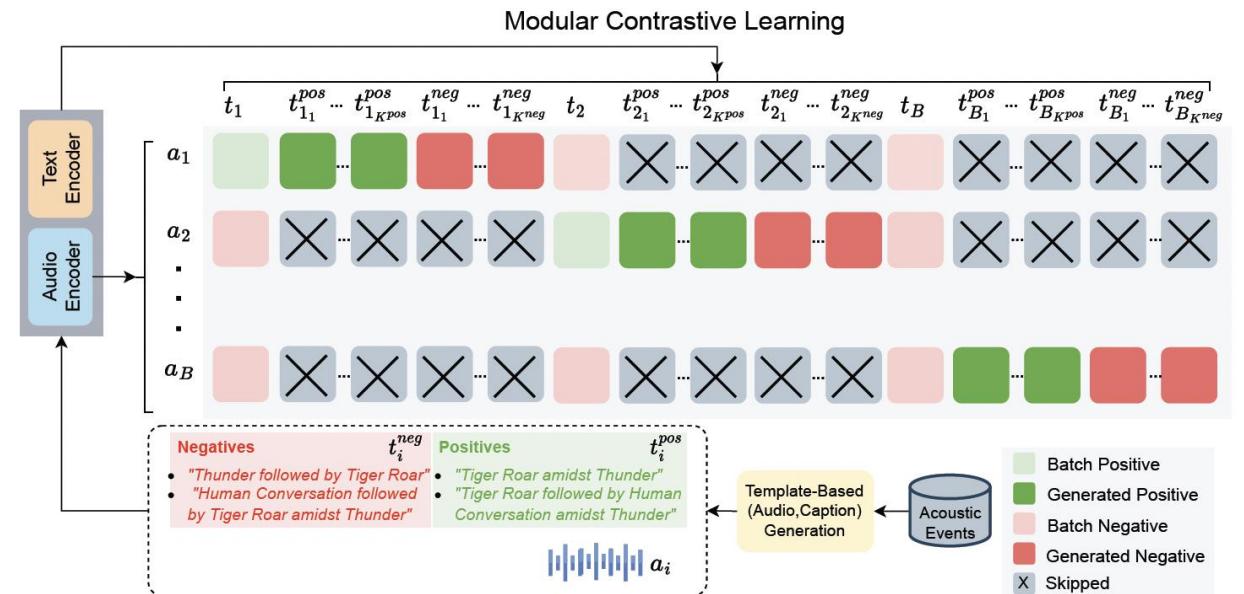
$$\ell_i^{a-2-t} = -a_i^\top t_i / \sigma + \log \left( \sum_{j=1}^B \exp(a_i^\top t_j / \sigma) + \sum_{k=1}^K \exp(a_i^\top t_{i_k}^\text{hard} / \sigma) \right)$$

# CompA: CompA-CLAP

**Modular Contrastive Learning:** We introduce a modular template-based approach that generates compositional audio-caption pairs from single acoustic events, aligning each audio with positives and compositionally-aware hard negatives. Using AudioSet snippets and LLM-generated scenes, we create high-quality synthetic data. Our training strategy optimizes contrastive loss to capture fine-grained compositional relationships



Synthetic data generation strategy.



$$\ell_i^{t-2-a} = - \left( \frac{1}{K^{pos}} \sum_{k=1}^{K^{pos}} (t_{i_k}^{pos})^\top a_i / \sigma \right) + \log \sum_{j=1}^B \exp(t_i^\top a_j / \sigma)$$

$$\ell_i^{a-2-t} = - \left( \frac{1}{K^{pos}} \sum_{k=1}^{K^{pos}} a_i^\top t_{i_k}^{pos} / \sigma \right) + \log \left( \sum_{j=1}^B \exp(a_i^\top t_j / \sigma) + \sum_{k=1}^{K^{neg}} \exp(a_i^\top t_{i_k}^{neg} / \sigma) \right)$$

# CompA: Results on Benchmark Datasets

Model	T-A Retrieval			A-T Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
MMT	36.1 / 6.7	72.0 / 21.6	84.5 / 33.2	39.6 / 7.0	76.8 / 22.7	86.7 / 34.6
ML-ACT	33.9 / 14.4	69.7 / 36.6	82.6 / 49.9	39.4 / 16.2	72.0 / 37.6	83.9 / 50.2
CLAP	34.6 / 16.7	70.2 / 41.1	82.0 / 54.1	41.9 / 20.0	73.1 / 44.9	84.6 / 58.7
CLAP-LAION	<b>36.2 / 17.2</b>	70.3 / 42.9	82.5 / 55.4	<u>45.0</u> / <b>24.2</b>	<u>76.7</u> / <u>51.1</u>	88.0 / 66.9
CLAP ( <i>ours</i> )	35.9 / 17.0	<b>78.3 / 44.1</b>	<u>89.6</u> / <b>56.9</b>	47.8 / 23.8	<u>83.2</u> / <b>51.8</b>	<b>90.7 / 67.8</b>
CompA-CLAP ( <i>ours</i> )	<u>36.1</u> / 16.8	<b>78.6 / 43.5</b>	<b>90.2 / 56.1</b>	<b>47.8</b> / 23.9	<b>83.5</b> / 50.7	90.2 / 67.6

Result comparison on AudioCaps/Clotho retrieval benchmarks.

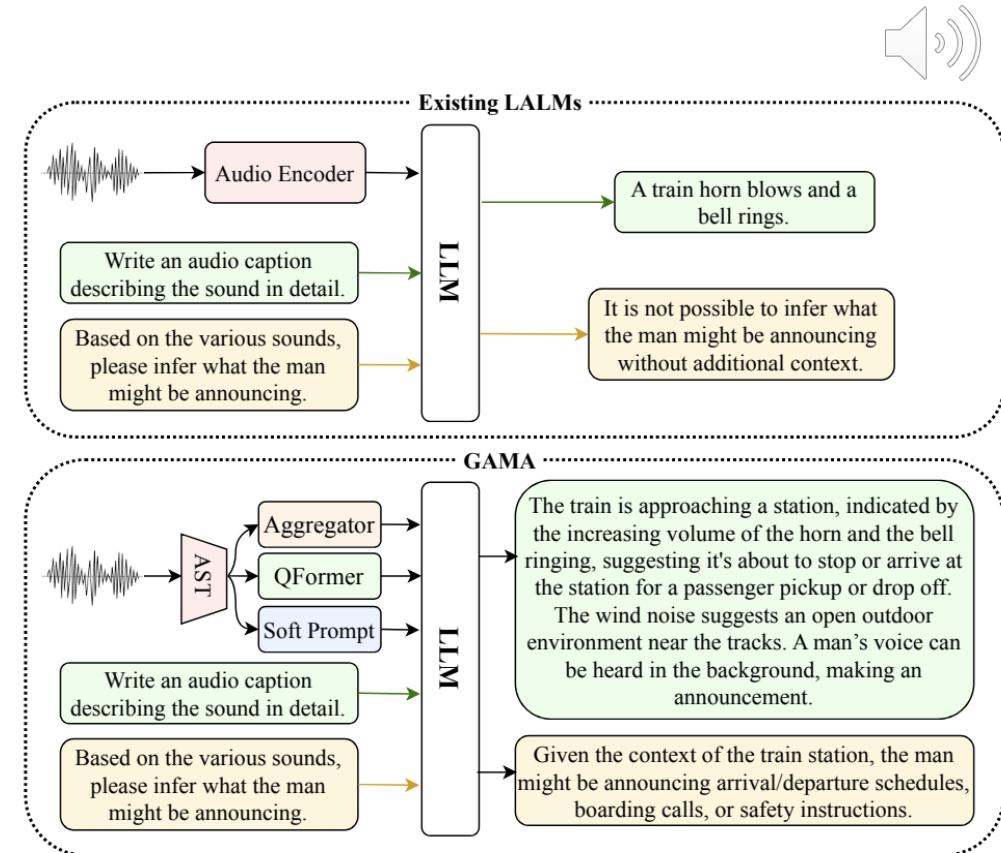
	ESC-50	US8K	VGGSound	FSD50K
Wav2CLIP	41.4	40.4	10.0	43.1
AudioClip	69.4	65.3	-	-
CLAP	82.6	73.2	-	58.6
CLAP-LAION-audio-630K	88.0	75.8	26.3	64.4
CLAP ( <i>ours</i> )	<b>90.2</b>	<b>86.1</b>	<u>29.1</u>	<b>77.8</b>
CompA-CLAP ( <i>ours</i> )	<u>89.1</u>	<u>85.7</u>	<b>29.5</b>	<u>77.4</u>

Result comparison on zero-shot audio classification benchmarks.



# GAMA: A General-purpose Large Audio-Language Model with Advanced Audio Understanding and Complex Reasoning Abilities (EMNLP 2024 Oral)

- Large Audio-Language Models (LALMs) struggle to respond to user queries that require complex reasoning.
- Most LALMs just use a single audio encoder with a linear layer to integrate audio representations to LLMs. This leads to limited audio understanding and increases hallucinations.
- GAMA is developed to improve both aspects!



Comparison of GAMA with LTU (Gong et al., 2023) on a query about an audio that involves complex reasoning.



# GAMA: Model Architecture

To improve audio perception, GAMA integrates multiple audio representations with a large language model to allow it to perceive diverse knowledge about an input audio.

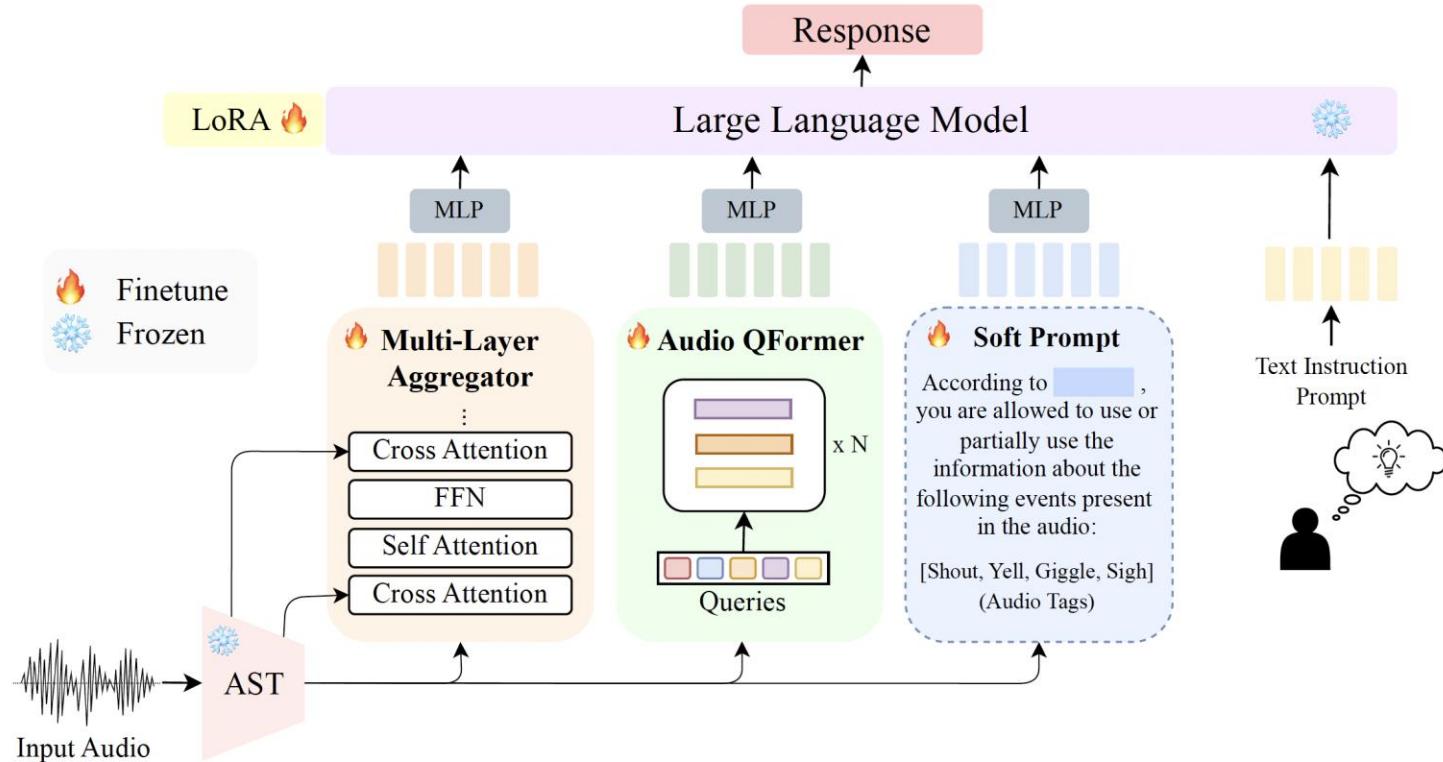
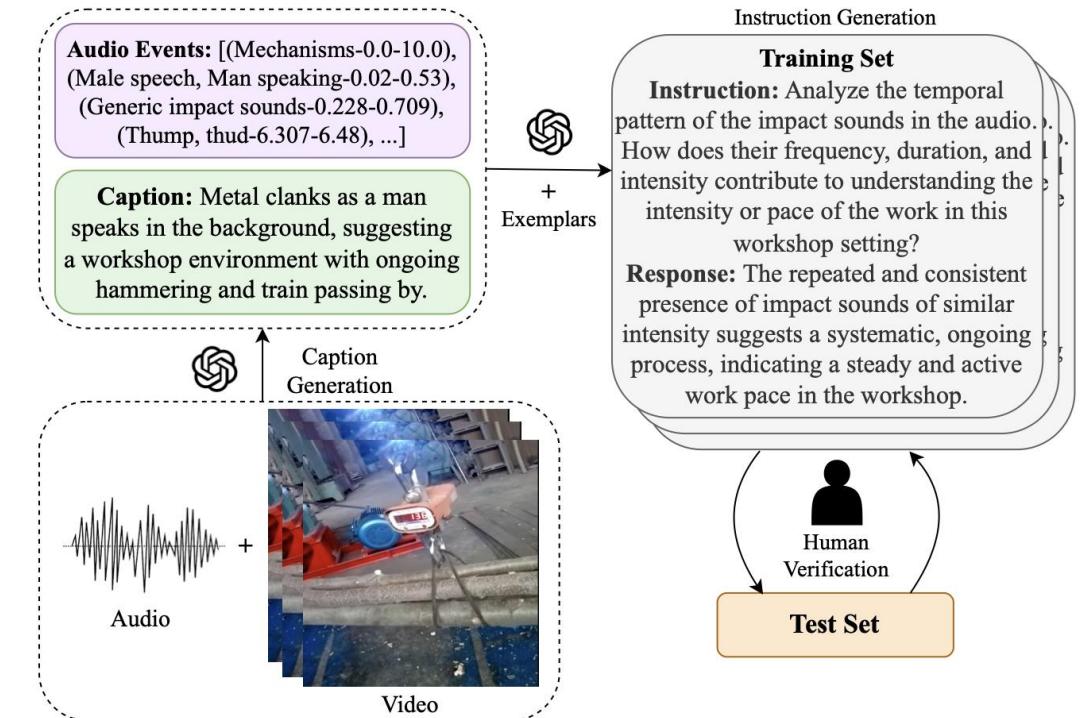


Illustration of the GAMA architecture.



# GAMA: CompA-R Reasoning Dataset

- GAMA is trained in 2 stages:
  - First, it is pre-trained on a large-scale dataset with audio-language pairs.
  - Next, it is instruction-tuned on CompA-R, (*Instruction-Tuning for Complex Audio Reasoning*) a dataset designed to train models to understand and reason about complex audio scenarios.



Our proposed pipeline to synthesize CompA-R.



# GAMA: Results on Benchmark Datasets

Model	ESC50# (Acc)	DCASE# (Mi-F1)	VS <sup>†</sup> (Acc)	TUT <sup>†</sup> (Acc)	BJO <sup>†</sup> (Acc)	VGG (mAP)	FSD (ACC)	NS <sub>ins.</sub> (ACC)	NS <sub>src.</sub> (ACC)	GTZAN <sup>†</sup> (ACC)	MSD <sup>†</sup> (ACC)	AudioSet (mAP)	Classif. Avg.	AudioCaps (SPICE)	Clotho (SPICE)	Cap. Avg.	ClothoAQA (ACC)
<i>Audio-Language encoder-based models. They are generalizable to unseen labels, but a pre-defined label set is required for inference.</i>																	
AudioCLIP	69.4	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
CLAP (2023a)	82.6	30.0	48.4	29.6	47.5	24.0	30.2	22.7	16.4	25.0	44.0	5.8	29.4	-	-	-	-
CLAP (2023a)	89.1	31.3	47.1	35.6	48.0	26.3	30.8	25.2	18.9	26.3	46.9	6.2	36.0	-	-	-	-
CompA-CLAP (2024b)	90.1	30.6	49.5	35.8	48.2	29.5	31.5	24.9	17.0	26.1	46.2	6.2	36.3	-	-	-	-
<i>Audio-Language generation-based models. They directly output label names and do not need a pre-defined label set is needed at inference.</i>																	
Qwen-Audio-Chat (2023)	71.7	32.4	74.2	16.9	50.8	17.5	39.8	30.2	41.3	41.6	69.1	13.4	41.1	14.7	9.8	12.3	32.3
LTU (2024)	81.7	37.5	<b>53.3</b>	19.9	67.8	50.3	43.9	28.0	41.8	9.9	74.2	18.3	42.4	16.9	11.7	<u>15.8</u>	25.1
SALMONN (2024)	16.4 <sup>†</sup>	18.0 <sup>†</sup>	16.9 <sup>†</sup>	7.8 <sup>†</sup>	25.0 <sup>†</sup>	23.3 <sup>†</sup>	22.1 <sup>†</sup>	16.2 <sup>†</sup>	33.7 <sup>†</sup>	10.1 <sup>†</sup>	28.8 <sup>†</sup>	13.4 <sup>†</sup>	17.9	8.3	7.6	8.0	23.1 <sup>†</sup>
Pengi (2023)	80.8 <sup>†</sup>	29.6 <sup>†</sup>	46.4 <sup>†</sup>	18.4 <sup>†</sup>	47.3 <sup>†</sup>	16.6 <sup>†</sup>	35.8	39.2	46.0	11.9	<b>93.0</b>	11.5	39.7	12.7	7.0	9.9	63.6
AudioGPT (2024)	41.3	20.9	35.8	14.9	21.6	5.6	18.8	40.9	15.6	11.9	28.5	12.7	22.4	6.9	6.2	6.6	33.4
<b>GAMA (ours)</b>	<b>82.6</b>	<b>38.4</b>	52.4	21.5	<b>69.5</b>	<b>52.2</b>	<b>47.8</b>	<b>63.9</b>	<b>99.5</b>	<b>13.8</b>	85.6	<b>19.2</b>	<b>53.9</b>	<b>18.5</b>	<b>13.5</b>	<b>16.0</b>	<b>71.6</b>
w/o AST & Aggregator	80.5	36.9	51.6	19.2	66.2	50.8	45.3	62.4	89.6	11.6	83.2	17.3	51.2	17.2	12.4	14.8	68.3
w/ Last Layer Features	81.3	37.6	50.2	20.4	68.2	51.7	45.8	<u>62.6</u>	92.3	11.2	81.5	18.1	51.7	<u>17.7</u>	12.8	15.3	<u>69.5</u>
w/o Audio Q-Former	79.7	37.4	51.3	20.2	68.0	51.6	46.4	<u>60.1</u>	90.4	11.6	79.8	18.4	51.2	<u>16.9</u>	11.9	14.4	<u>61.2</u>
w/ CLAP	81.8	38.4	52.2	<b>21.6</b>	<b>69.1</b>	<b>52.0</b>	<b>47.5</b>	58.8	99.5	<u>12.4</u>	77.9	<u>19.0</u>	<u>52.5</u>	17.2	<u>13.1</u>	15.1	66.4

Comparison of GAMA with baselines on evaluation datasets described on close-ended general audio and music understanding benchmarks.



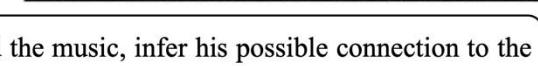
# GAMA: Results on Reasoning Datasets

Models	CompA-R-test (GPT-4/Human)				OpenAQA				Dense Captioning		
	Clarity	Correctness	Engagement	Avg.	Clarity	Correctness	Engagement	Avg.	AudioCaps	Clotho	Avg.
Qwen-Audio-Chat (2023)	3.5 / 3.4	3.3 / 3.4	3.6 / 3.7	3.5 / 3.5	3.6	3.6	3.5	3.6	3.8	3.6	3.7
LTU (2024)	3.5 / 4.0	3.2 / 3.3	3.4 / 3.5	3.4 / 3.6	3.5	3.7	3.5	3.6	3.5	3.6	3.5
SALMONN (2024)	2.6 / 2.8	2.4 / 2.3	2.0 / 2.2	2.3 / 2.4	2.4	2.5	2.7	2.5	2.8	3.1	2.9
Pengi (2023)	1.8 / 1.6	1.5 / 1.4	1.3 / 1.2	1.5 / 1.4	1.7	1.5	1.4	1.5	2.6	2.8	2.7
AudioGPT (2024)	1.3 / 1.4	1.6 / 1.5	1.4 / 1.7	1.4 / 1.5	1.6	1.5	1.5	1.5	2.7	2.9	2.8
LTU w/ CompA-R ( <i>ours</i> )	3.5 / 4.0	3.2 / 3.3	3.4 / 3.5	3.6 / 3.6	3.5	3.7	3.5	3.6	3.7	3.8	3.8
<b>GAMA-IT (<i>ours</i>)</b>	<b>4.3 / 4.5</b>	<b>3.9 / 4.1</b>	<b>3.9 / 4.3</b>	<b>4.0 / 4.3</b>	<b>4.0</b>	<b>4.2</b>	<b>3.8</b>	<b>4.0</b>	<b>4.3</b>	<b>4.1</b>	<b>4.2</b>
w/o Soft Prompt	4.1 / 4.2	3.7 / 3.8	3.6 / 3.4	3.8 / 3.8	3.9	3.8	3.7	3.8	4.1	3.9	4.0
w/o Aggregator	4.0 / 4.2	3.5 / 3.5	3.6 / 3.5	3.7 / 3.7	3.7	3.7	3.5	3.6	3.7	3.8	3.8
w/o Audio Q-Former	3.8 / 3.7	3.4 / 3.6	3.5 / 3.3	3.6 / 3.5	3.4	3.9	3.5	3.6	3.7	3.5	3.6

Comparison of GAMA with other baselines on open-ended AQA on OpenAQA, complex open-ended AQA on CompA-R- test and Dense Captioning on 500 instances from AudioCaps and Clotho.



# GAMA: Qualitative Examples



**Question:** Based on the given audio, infer the type of environment or setting. Consider the presence and duration of car sounds.

LTU

The audio clip suggests an urban or suburban environment with a mix of residential and commercial areas. The presence of car sounds suggests that the environment is a busy one.

HE : 3

SALMNON

The given audio suggests that the environment or setting is a busy city street with a lot of car traffic. The audio also indicates that the car sounds are continuous and long-lasting. Therefore, it can be inferred that the audio is from a busy city street with a lot of car traffic.

HE : 3

GAMA-IT Pengi

It is gasoline.

HE : 2

The environment is likely a race track.

HE : 5

LTU

The male speaker's speech adds a sense of authority and control to the event's atmosphere, as well as a sense of urgency or importance to the crowd's reactions. The speech also provides a contrast to the music and crowd reactions, creating a dynamic and engaging soundscape that draws the listener's attention and emotions.

HE : 2

SALMNON

The male speaker is interspersed with the music and crowd reactions to create a dynamic and engaging atmosphere at the event. His speech adds energy and excitement to the event, keeps the audience engaged and involved, and helps create a memorable and enjoyable experience for all attendees.

HE : 1

GAMA-IT Pengi

It is music.

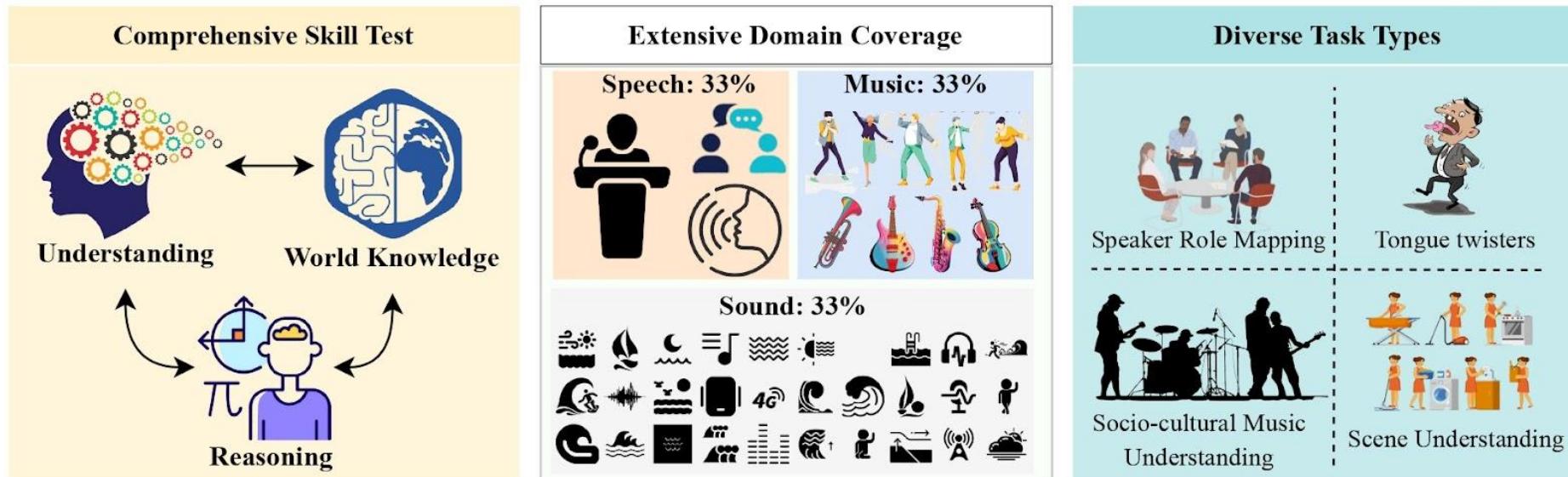
HE : 1

The man is likely a guitarist or an instructor who is explaining the process of tuning a guitar

HE : 5

# MMAU: A Massive Multi-Task Audio Understanding and Reasoning Benchmark (ICLR 2025 Spotlight)

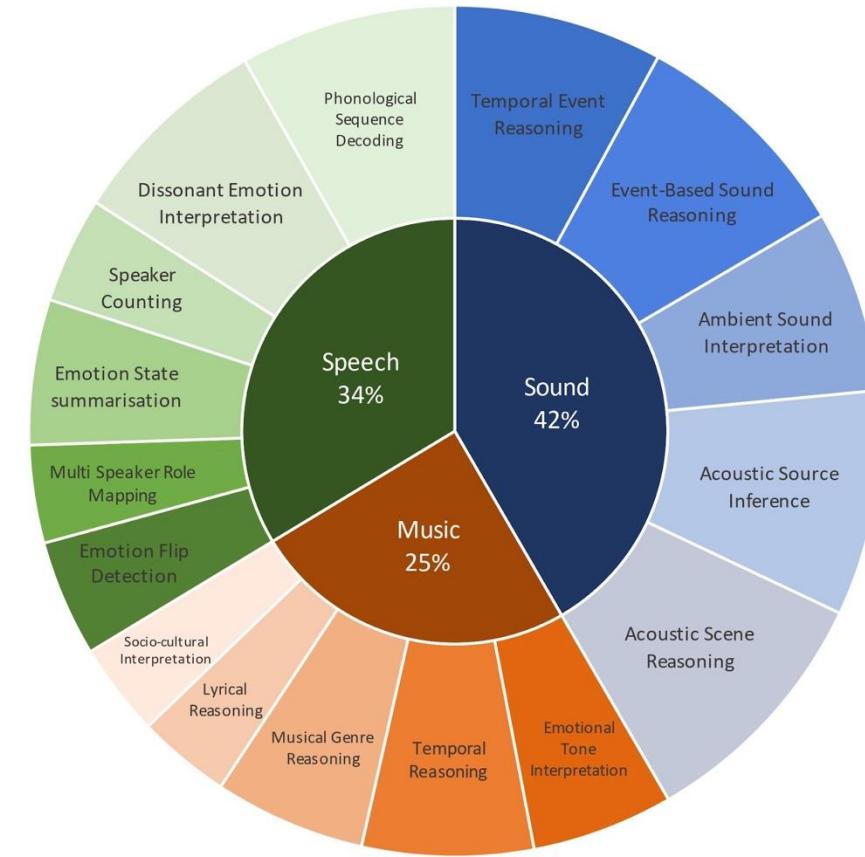
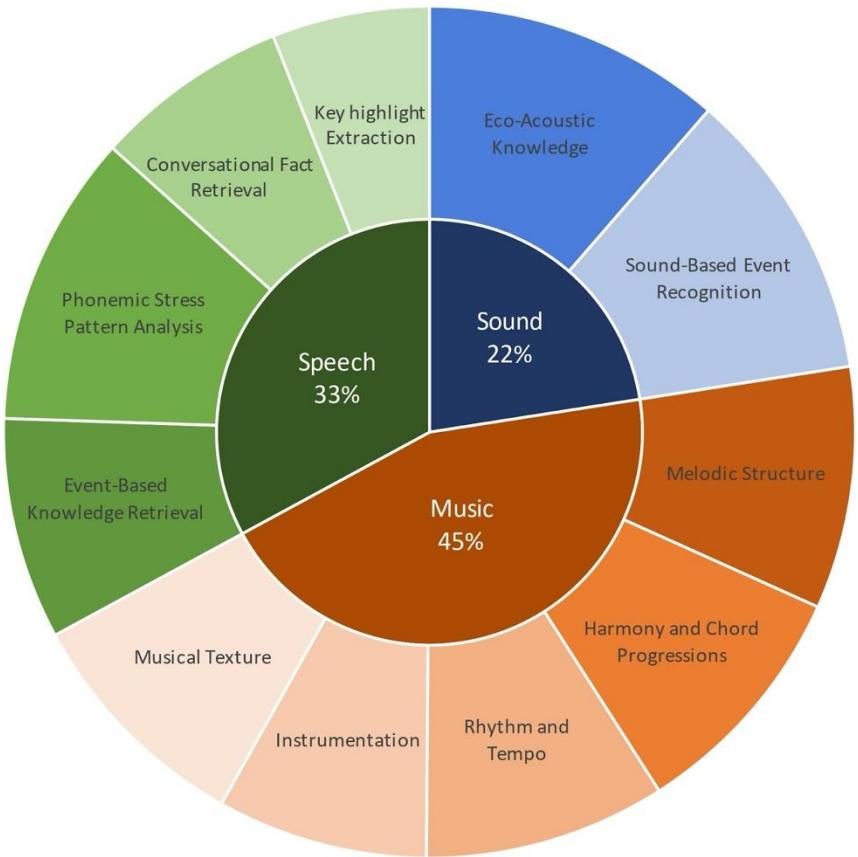
- [Morris et al., 2024] **Artificial General Intelligence (AGI)** as a system that performs at the “90th percentile of skilled adults” across a wide array of tasks.
- Current audio benchmarks **fall short of this standard**. Tasks such as speech recognition do not demand the expertise of skilled humans and can often be performed by young children (Lippmann, 1997; Gerhardstein & Rovee-Collier, 2002)
- MMLU (for language) and MMMU (for vision) have pushed the boundaries of model capabilities, prompting incremental improvements. Nothing as such exists in audio.



We present MMAU, the most comprehensive audio understanding and reasoning benchmark. MMAU comprises 10k carefully curated audio clips paired with natural language questions and answers spanning speech, environmental sounds, and music.

# MMAU: Skill Distribution

MMAU is composed of questions that challenge models no 27 distinct skills, **16 Reasoning (right)** and **11 Information Extraction Tasks (left)**.

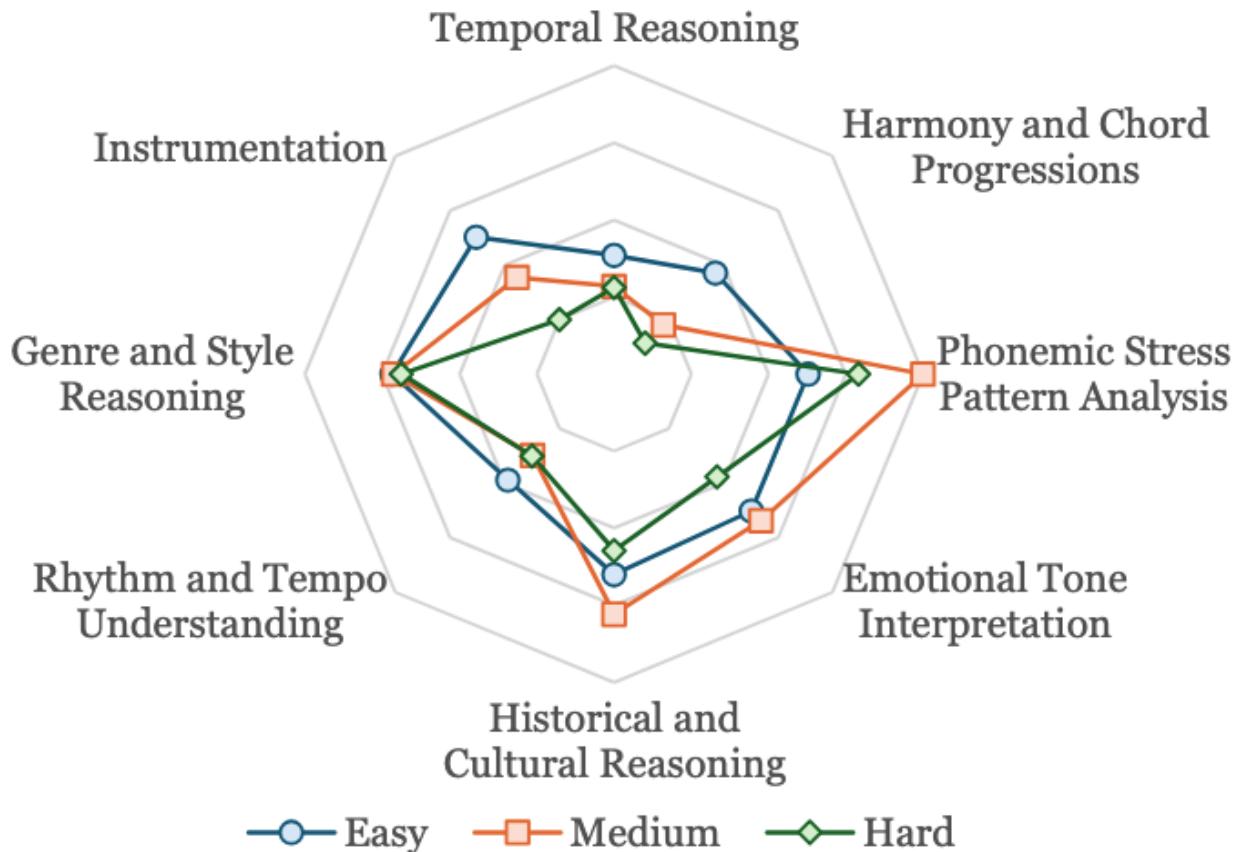


# MMAU: Scores for Frontier Models

Models	Size	{So, Mu, Sp}	Sound		Music		Speech		Avg	
			Test-mini	Test	Test-mini	Test	Test-mini	Test	Test-mini	Test
Random Guess	-	-	26.72	25.73	24.55	26.53	26.72	25.50	26.00	25.92
Most Frequent Choice	-	-	27.02	25.73	20.35	23.73	29.12	30.33	25.50	26.50
Human (test-mini)	-	-	86.31	-	78.22	-	82.17	-	82.23	-
<b>Large Audio Language Models (LALMs)</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
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	✓ ✗ ×	55.25	<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>	42.04	45.90	49.20	52.50
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>	54.46	49.40	48.56	<b>58.55</b>	<b>55.90</b>	<b>54.90</b>	<b>52.97</b>
<b>Large Language Models (LLMs)</b>										
GPT4o + weak cap.	-	-	39.33	35.80	39.52	41.9	<b>58.25</b>	68.27	45.70	48.65
GPT4o + strong cap.	-	-	<b>57.35</b>	<b>55.83</b>	49.70	<b>51.73</b>	<b>64.86</b>	<b>68.66</b>	<b>57.30</b>	<b>58.74</b>
Llama-3-Ins. + weak cap.	8B	-	34.23	33.73	38.02	42.36	54.05	61.54	42.10	45.87
Llama-3-Ins. + strong cap.	8B	-	50.75	49.10	<b>50.29</b>	48.93	55.25	62.70	<b>52.10</b>	<b>53.57</b>

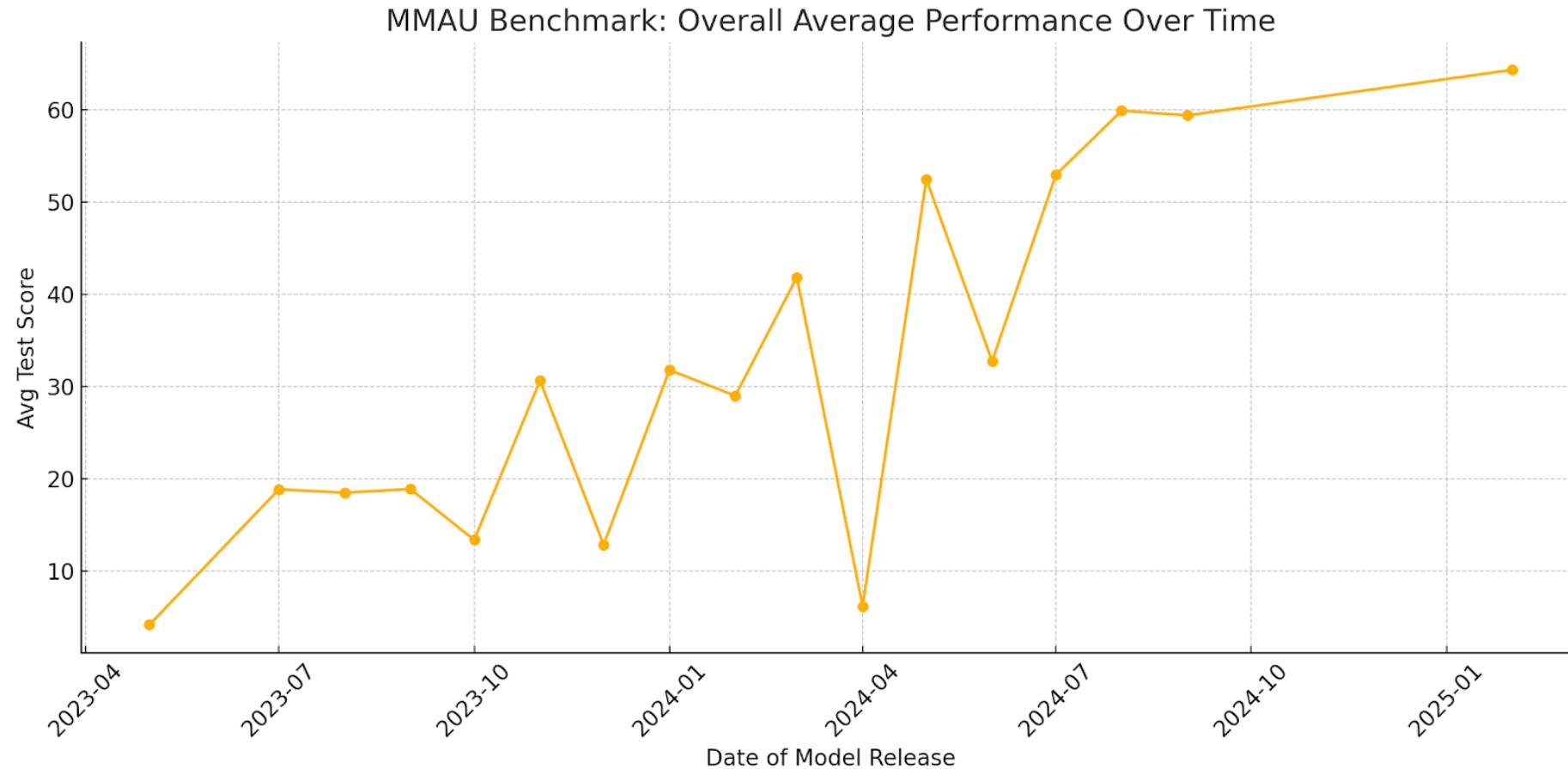
Most models perform poorly on MMAU, thereby highlighting significant gap.

# MMAU: Skill-Specific Performance



Accuracy distribution for Gemini 2.0 Flash across easy, medium, and hard questions, categorized by skill type. The graph highlights how LALMs excel in some skills across all difficulty levels (e.g., Phonemic Stress Pattern Analysis) but struggle with others (e.g., Temporal Reasoning) regardless of difficulty.

# MMAU: Model Performance Across Time

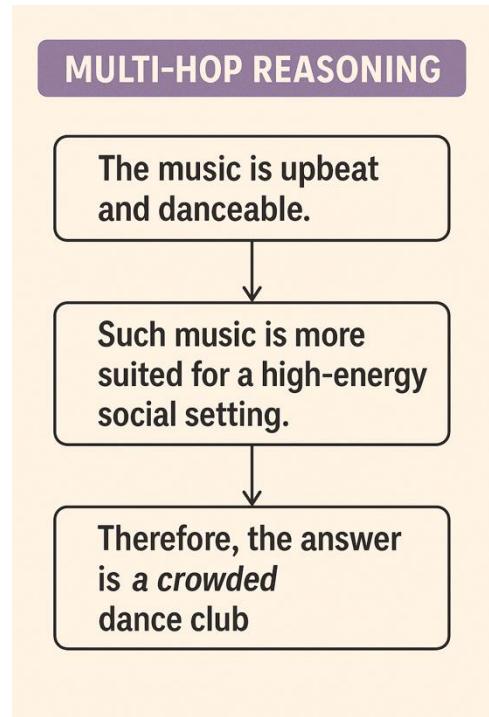


# MMAU-Pro: Complex Reasoning Benchmark (Work in progress)

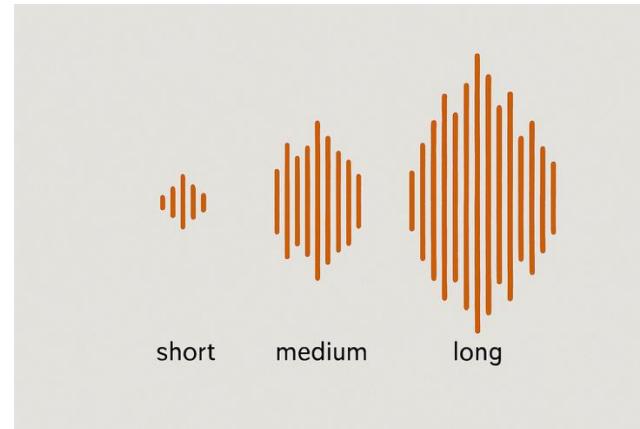
MMAU-Pro extends the original MMAU benchmark by incorporating more complex evaluation settings, assessing model across more challenging skills. Some unique features of MMAU-Pro are:



Multi-cultural music and sound evaluation.



Questions requiring multi-hop reasoning.  
E.g., *In what setting would this music most likely be heard?*



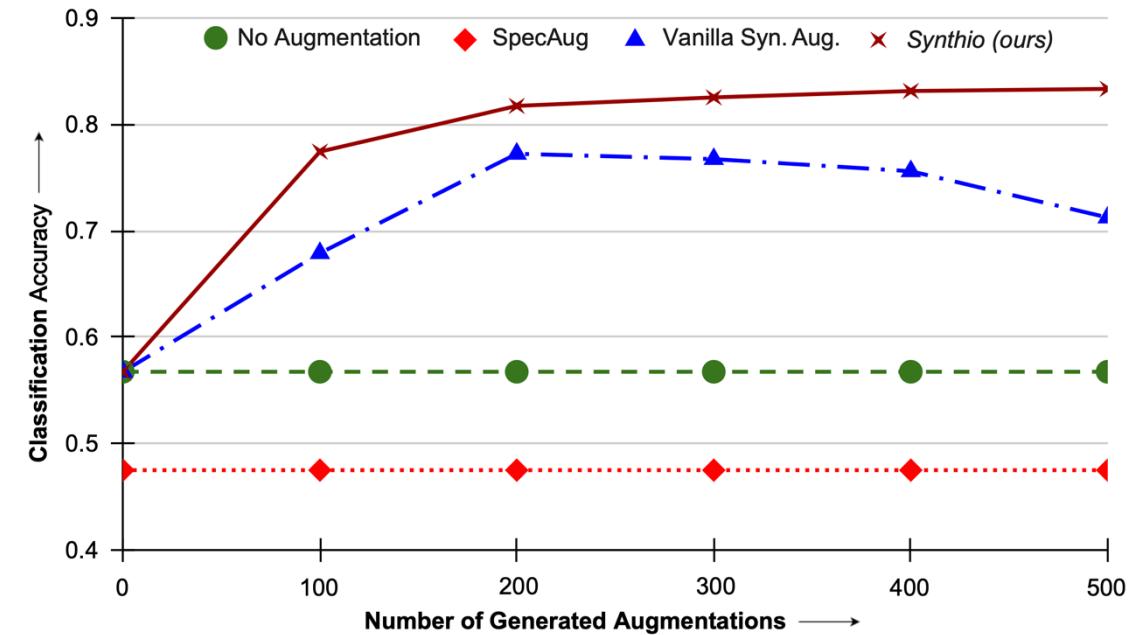
Variable length including long audio evaluation.



Assess skills relevant to speech-to-speech interaction.

# Synthio: Augmenting Small-Scale Audio Classification Datasets with Synthetic Data (ICLR 2025)

- The use of synthetic data to improve understanding and reasoning is booming in language and vision (e.g., Phi4, GPT-4o). However, in audio, there is lack of foundational audio generation models.
- Existing data augmentation methods in audio (noise augmentation or masking) do not capture the true diversity present in real-world audios.
- We propose **Synthio**, a novel synthetic data generation technique to generate diverse and consistent synthetic audios to improve audio classification.



Traditional augmentation degrades performance on small-scale datasets. *Synthio further enhances performance by generating consistent and diverse synthetic data.*

# Synthio: Motivation

**Problem 1 – Alignment:** When prompted with “sound of a *bus*” for the category *bus* in the TUT-Urban dataset, the generated audio may not reflect the typical bus sounds in European cities (where TUT was recorded), as bus sounds can vary by region, with some featuring loud engines and dense crowds while others have quieter engines and sparse crowds.



Sound of a bus from the TUT Urban dataset.



Sound of a bus from a text-to-audio model  
trained on AudioSet.

Can we just fine-tune the text-to-audio model? -- Fine-tuning on small datasets lead to overfitting and generation of audios that sound similar.

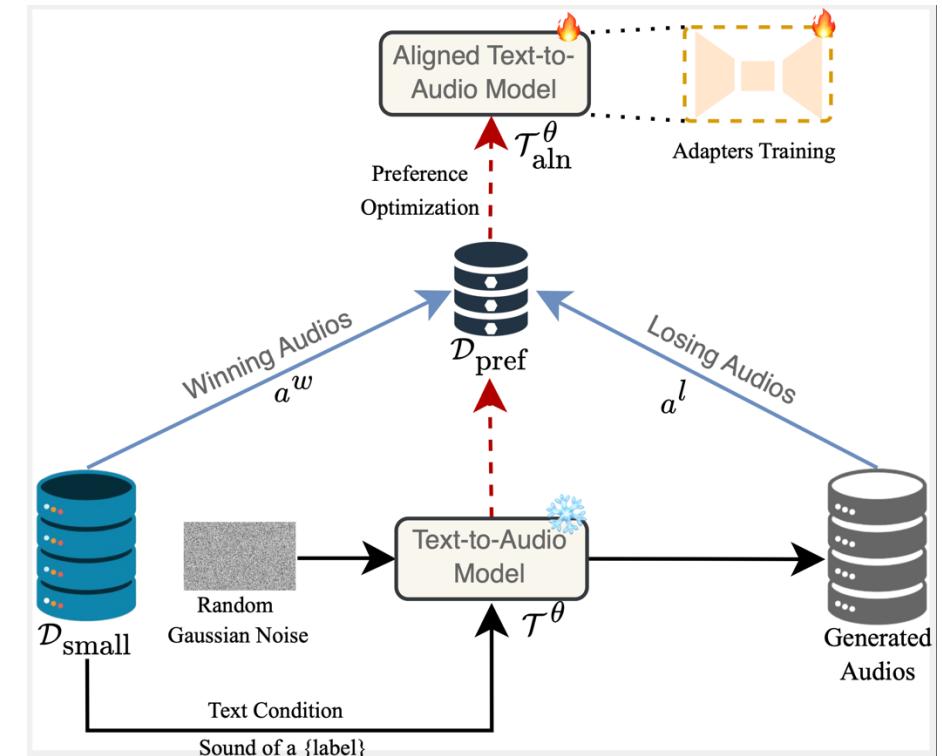
**Problem 2 – Lack of feature diversity:** When prompted with “Sound of a park”, we observed that 9 out of 10 times, the model generated the sound of children playing as part of the generated audio.



Two sounds with different random seeds,  
CFG and inference steps.

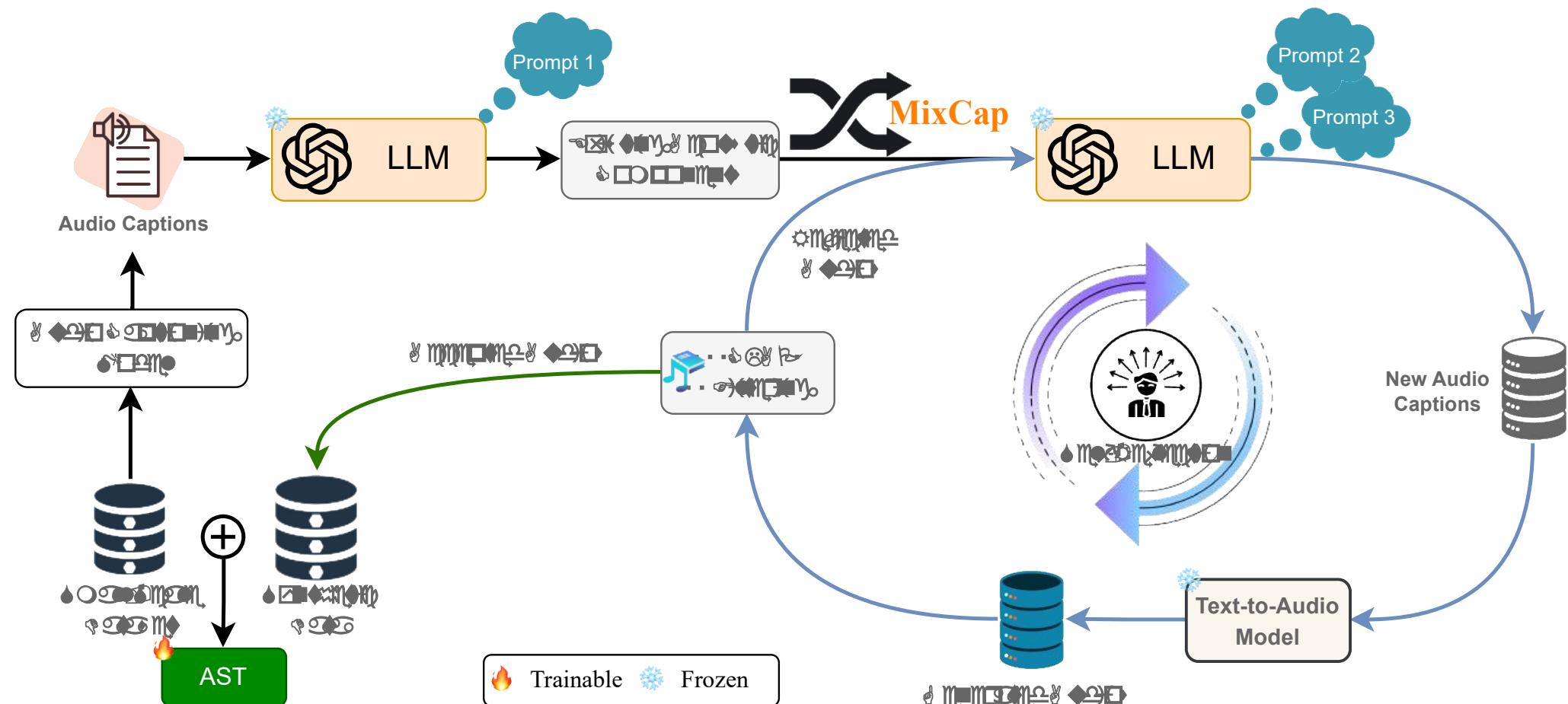
# Synthio: Aligning the Text-to-Audio Model using Preference Optimization

- Step 1: Constructing the Preference Dataset
  - Generate template-based captions for each instance (e.g., "Sound of a [label]").
  - Prompt the T2A model multiple times per instance, pairing generated audio with ground-truth audio, treating generated as "loser" and ground-truth as "winner."
- Step 2: Preference Optimization with DPO (Rafailov et. al)
  - Train the T2A model on the preference dataset using DPO.
  - The optimized model improves alignment with ground-truth audio.



We align the T2A model with the small-scale dataset using DPO, ensuring generated audio matches its acoustic characteristics.

# Synthio: Generating Diverse Synthetic Augmentations



We align the T2A model with the small-scale dataset using DPO, ensuring generated audio matches its acoustic characteristics.

# Synthio: Results

<i>n</i>	Method	ESC-50	USD8K	GTZAN	Medley	TUT	NSynth	VS	MSDB	DCASE	FSD50K
50	Gold-only (No Aug.)	22.25	55.09	47.05	47.23	37.60	33.32	77.49	56.85	12.09	8.69
	Random Noise	18.50	57.42	45.20	46.55	35.86	32.42	76.41	52.55	13.21	9.15
	Pitch Shifting	20.55	59.32	46.80	48.17	37.22	34.34	78.17	54.50	12.93	10.04
	SpecAugment	19.50	58.36	46.00	47.18	36.73	27.32	77.27	53.25	12.81	9.22
	Audiomentations	20.35	60.13	47.25	48.30	38.24	28.15	79.12	54.51	13.28	11.44
	Retrieval	19.20	37.14	42.55	43.65	35.80	31.27	71.42	51.35	10.53	7.28
	Vanilla Syn. Aug.	40.75	63.54	55.35	47.23	41.50	33.17	78.37	54.10	15.89	11.76
	+ LLM Caps.	36.80	65.84	63.74	55.36	40.90	38.17	78.77	57.05	13.07	11.84
	<b>Synthio (ours)</b>	<b>49.50<sub>+122%</sub></b>	<b>76.12<sub>+38%</sub></b>	<b>68.20<sub>+44%</sub></b>	<b>60.58<sub>+28%</sub></b>	<b>43.84<sub>+17%</sub></b>	<b>40.83<sub>+22%</sub></b>	<b>80.67<sub>+4%</sub></b>	<b>60.15<sub>+5%</sub></b>	<b>17.23<sub>+42%</sub></b>	<b>15.41<sub>+77%</sub></b>
	w/ Template Captions	41.25	66.11	64.40	54.52	41.37	37.52	78.57	59.60	14.15	13.6
100	w/ ERM	41.30	69.80	61.70	56.60	42.00	38.62	79.75	57.75	13.28	14.17
	w/o Self-Reflection	45.25	72.57	64.55	58.00	42.81	39.50	78.56	57.25	15.63	14.02
	w/o MixCap	42.70	64.72	54.65	52.18	41.93	36.13	78.70	58.80	14.82	13.61
	w/o DPO	36.55	68.12	56.10	52.55	41.39	40.31	79.03	57.55	14.53	11.28
	Gold-only (No Aug.)	56.75	72.89	64.15	57.81	47.14	39.11	84.32	65.60	12.50	10.70
	Random Noise	58.50	71.54	65.50	56.98	46.21	38.20	83.33	66.15	13.35	14.95
	Pitch Shifting	59.55	73.52	66.75	58.46	47.50	39.53	85.07	68.25	12.19	15.02
	SpecAugment	47.50	72.43	69.75	58.06	50.07	41.96	85.14	66.40	14.17	15.72
	Audiomentations	48.50	73.82	71.05	59.32	51.14	42.15	85.24	68.40	16.93	14.29
	Retrieval	52.45	68.24	61.55	54.83	45.39	37.84	83.27	58.55	10.93	10.05
200	Vanilla Syn. Aug.	77.25	77.31	68.25	63.58	49.96	42.31	84.78	63.55	15.73	13.78
	+ LLM Caps.	67.05	79.73	67.90	65.79	48.63	41.83	84.83	65.95	16.32	14.12
	<b>Synthio (ours)</b>	<b>83.35<sub>+47%</sub></b>	<b>85.00<sub>+17%</sub></b>	<b>71.20<sub>+11%</sub></b>	<b>71.23<sub>+23%</sub></b>	<b>52.42<sub>+11%</sub></b>	<b>44.92<sub>+15%</sub></b>	<b>86.70<sub>+3%</sub></b>	<b>68.80<sub>+5%</sub></b>	<b>19.38<sub>+55%</sub></b>	<b>17.33<sub>+62%</sub></b>
	w/ Template Captions	78.00	80.32	68.15	64.20	49.95	42.76	85.11	66.05	16.32	14.32
	w/ ERM	73.20	81.81	67.25	66.57	51.11	43.74	84.73	68.00	17.21	15.19
400	w/o Self-Reflection	77.65	82.38	69.55	68.52	51.75	44.38	82.53	66.20	15.89	13.96
	w/o MixCap	73.50	78.30	68.50	66.52	50.63	42.27	83.52	66.35	16.77	14.93
	w/o DPO	66.75	75.46	66.15	60.81	48.78	40.31	84.67	67.85	14.83	13.92

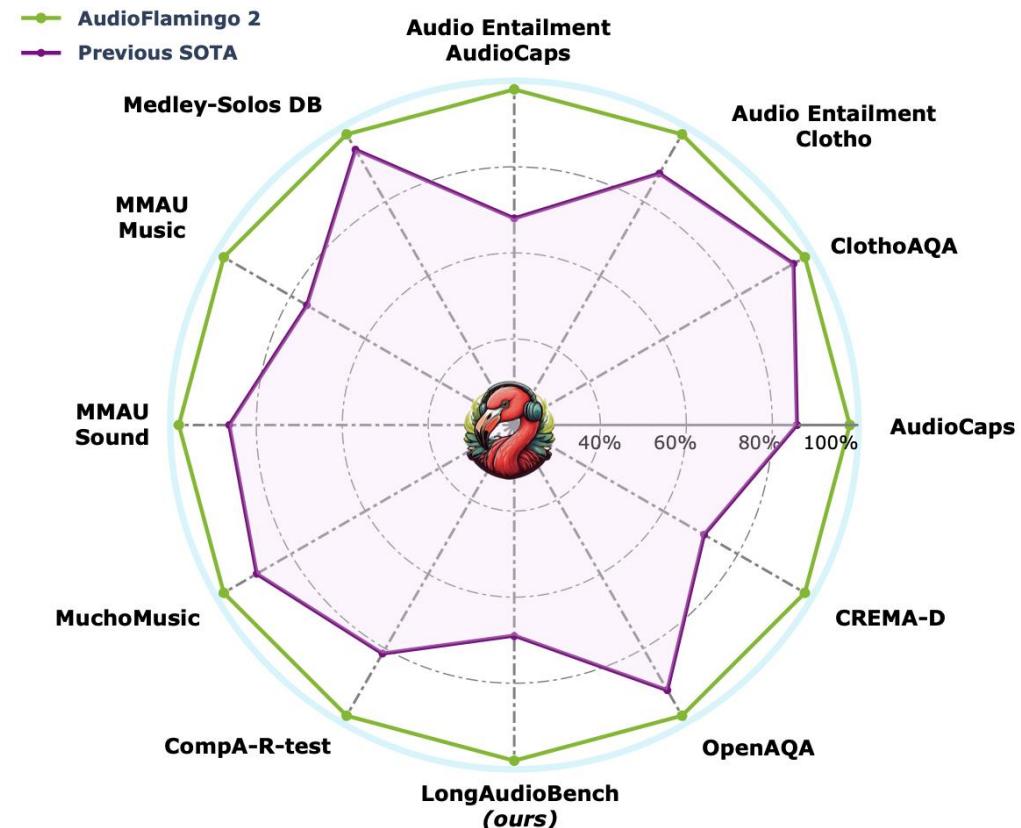
Results on benchmark datasets with just 50 and 100 gold training samples. Synthio shows huge performance gains compared to our baselines.



# Audio Flamingo 2: An Audio-Language Model with Long-Audio Understanding and Expert Reasoning Abilities (Joint with NVIDIA, Feb 2025)

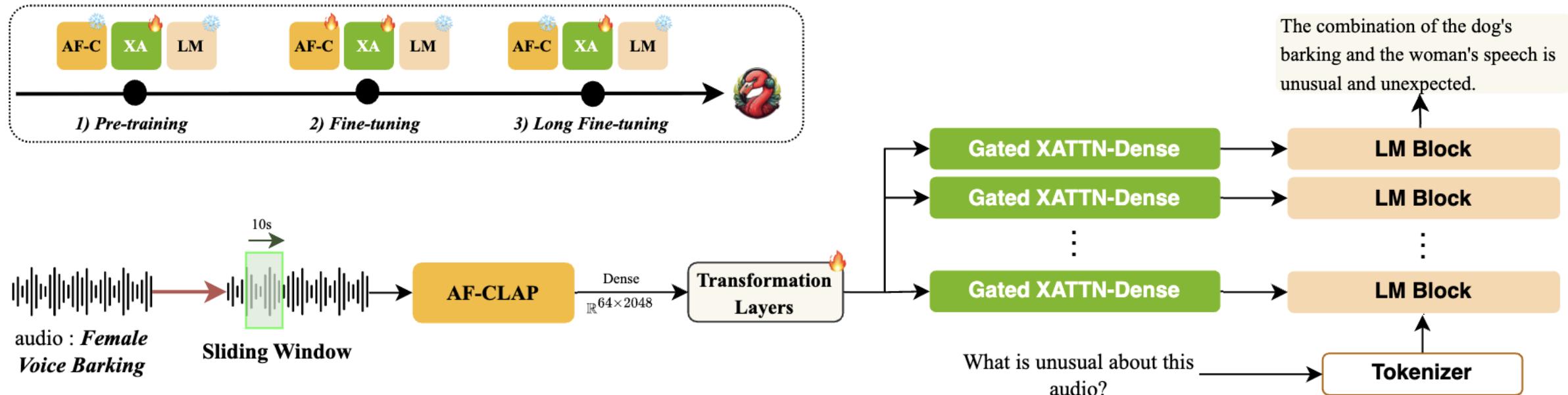
Audio Flamingo 2 is a SOTA LALM that focuses on 2 main aspects:

- Improving audio understanding and reasoning
  - We develop **AF-CLAP**, a custom audio encoder for robust representations
  - Generate synthetic reasoning data at scale
- Introduce Long Audio Understanding (30 seconds – 5 minutes)
  - We introduce **LongAudio**, the first long audio reasoning dataset
  - We propose several novel training strategies



Audio Flamingo 2 (AF2) versus previous SOTA on a number of audio understanding and reasoning benchmarks. **AF2 outperformed all models while being the smallest.**

# AF2: Overall Architecture



Audio Flamingo 2 is based on the cross-attention architecture, where we condition our custom CLAP.  
The model is trained on 3 stages of training and can perceive audios up to 5 minutes in length.

# AF2: Improved Audio Representations with AF-CLAP

We focus on two primary aspects in AF-CLAP :

1. We scale training data to 8M audio-caption pairs:
  - i. Generate new data from long videos, inspired from [Venkataramanan et al., 2024]
2. We propose a new training objective
  - i. Improve linguistic invariance
  - ii. Improve compositional reasoning

**Original Caption:** A dog barking followed by the sound of a train approaching.

**Positive:** A dog barking followed by the sound of a railcar approaching.

**Negative:** A dog barking preceded by the sound of a railcar approaching.

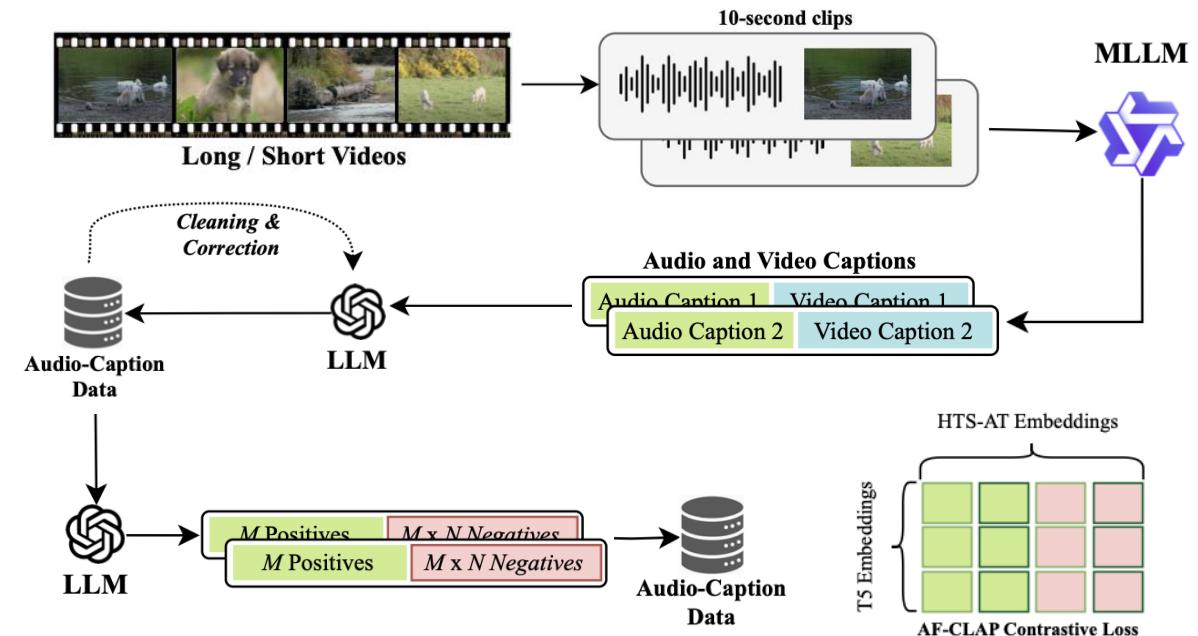
- iii. Improve the CLAP training objective

$$\mathcal{L} = -\frac{1}{B} \sum_{i=1}^B \log \frac{S(i, i)}{S_{\text{neg}}(i) + \sum_{j=1}^B S(j, i)},$$

where

$$S(j, i) = \sum_m s(\mathcal{T}(\mathcal{P}(x_j)_m), \mathcal{A}(x_i)),$$

$$S_{\text{neg}}(i) = \sum_{m,n} s(\mathcal{T}(\mathcal{N}(x_i)_{m,n}), \mathcal{A}(x_i)),$$



Overall AF-CLAP training pipeline.

# AF2: AudioSkills – Large-Scale Synthetic Reasoning Data

We propose AudioSkills, the 1<sup>st</sup> and the largest audio reasoning data with **~4.2M AQA pairs**. AudioSkills comprises of skill-wise reasoning data with seven distinct skills.

## Other AQA Datasets

**Foundational:** What is the genre of the music?

**Answer:** Rock

**ClothoAQA:** Are there people having conversation?

**Answer:** Yes

**OpenAQA:** What kind of sound do the mechanisms make?

**Answer:** Mechanical sounds

**SALMONN:** What is happening to the vehicle?

**Answer:** Vehicle is accelerating.

## Temporal Relationship Identification

**Order:** In what sequence do the sounds first appear in the audio?  
(A) Rain (B) Human voice ....

**Attribute:** How does the sound of thunder change over time?  
(A) Gets louder (B) Gets softer ....

**Grounding:** When does the human voice appear in the audio?  
(A) Beginning (B) Middle (C) End

**Referring:** What sound appears last in the audio?  
(A) Rain (B) Human voice ....

## General Reasoning

**Question:** How does the melody in the audio contribute to the hypnotic effect of the music?  
(A) By changing frequently to maintain interest. (B) By maintaining a consistent and repetitive loop ....

## Attribute Identification

**Question:** Which is the loudest event in the audio?  
**Answer:** The loudest event in the audio is the sound of a dog barking.

## Counting

**Level 1 & 2:** How many times did the dog bark?

**Answer:** Two

**Level 3:** How many times did the second sound occur in the entire audio?

**Answer:** Four

## Information Extraction

**Question:** Which instrument primarily provides the bass line that complements the female vocalist's melody?  
(A) Piano (B) Tuba (C) Guitar ...

## Contextual Speech Event Reasoning

**Question:** How does the repetition in the spoken utterance influence the interpretation of the sequence of actions taking place?

**Answer:** The repetition suggests urgency or frustration in an effort to manage an ongoing situation, possibly highlighted by background sounds of movement.

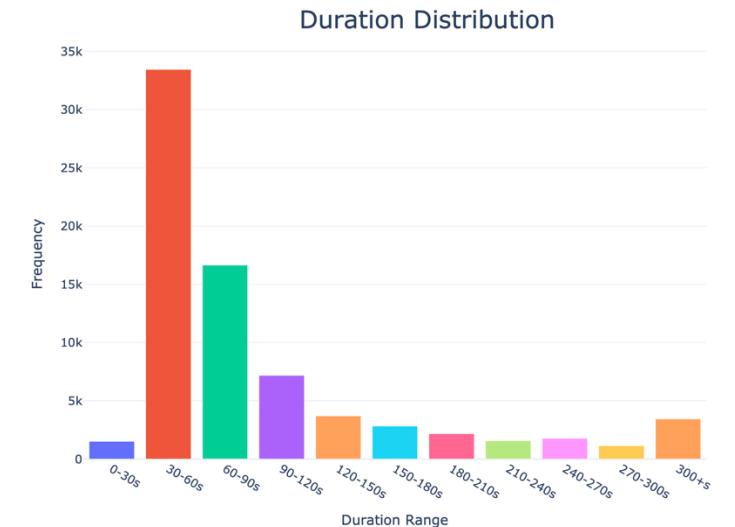
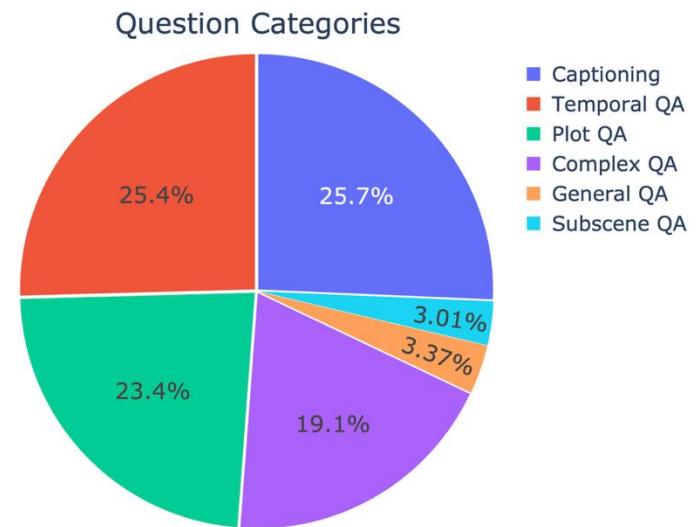
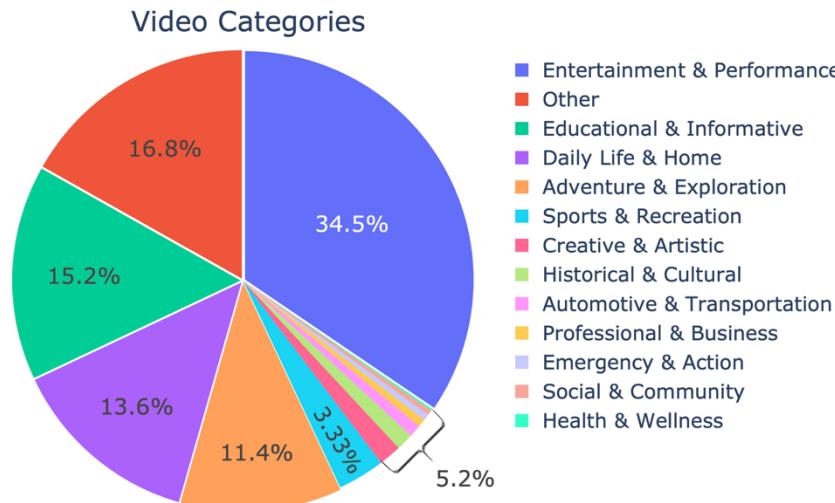
## Contextual Sound Event Reasoning

**Question:** Considering the correlation of the man's speech and the music, infer his possible connection to the music playing in the scene.

**Answer:** The man is likely a guitarist or an instructor who is explaining the process of tuning a guitar.

# AF2: LongAudio – A Long Audio Understanding Dataset

The proportion of video categories and distribution of durations for the **LongAudio** dataset with **262,928** unique AQA and **80k** unique audios. We target captioning and reasoning tasks.



# AF2-3B: Results on Benchmark Datasets

Dataset	Task	Previous SOTA	Ours	Dataset	Task	Previous SOTA	Ours
ClothoAQAnan.	AQA - ACC	74.9% - Qwen2-A	<b>86.9%</b> +12.0%	Clotho-v2	CAP - CIDEr	0.45 - Qwen2-A	<b>0.46</b> +0.01
ClothoAQAnon-bin	AQA - ACC	49.5% - AF	<b>52.6%</b> +3.1%	AudioCaps <sup>ZS</sup>	CAP - CIDEr	0.46 - AF	<b>0.58</b> +0.12
MusicAVQAAudio	AQA - ACC	72.1% - Qwen-A	<b>72.3%</b> +0.2%	CREMA-D <sup>ZS</sup>	CLS - ACC	26.5% - AF	<b>36.6%</b> +10.1%
NonSpeech7k	CLS - ACC	83.9% - AF	<b>84.3%</b> +0.4%	Ravdess <sup>ZS</sup>	CLS - ACC	20.9% - AF	<b>26.3%</b> +5.4%
CochlScene	CLS - ACC	<b>91.6%</b> - Pengi	82.1% -9.5%	GTZAN <sup>ZS</sup>	CLS - ACC	65.2% - AF	<b>69.1%</b> +3.9%
NS <sub>source</sub>	CLS - ACC	60.1% - Pengi	<b>62.0%</b> +1.9%	Medley-solos-DB <sup>ZS</sup>	CLS - ACC	85.6% - GAMA	<b>85.8%</b> +0.2%
NS <sub>instrument</sub>	CLS - ACC	<b>78.8%</b> - Qwen-A	71.13% -7.67%	US8K <sup>ZS</sup>	CLS - ACC	<b>71.2%</b> - AF	68.0% -3.2%
FSD50k	CLS - mAP	47.9% - GAMA	<b>49.2%</b> +1.3%	ESC50 <sup>ZS</sup>	CLS - ACC	80.6% - GAMA	<b>83.9%</b> +3.3%

Results on Audio Classification and Captioning Tasks. AF2 outperforms larger models on most of these tasks.

# AF2-3B : Results on Reasoning Datasets

Dataset	Previous SOTA	Ours
MMAU Sound	61.7% - Gemini F v2	<b>65.1%</b> +3.4%
MMAU Music	56.5% - Gemini F v2	<b>72.9%</b> +16.4%
AE Clotho	83.3% - Qwen-A	<b>92.5%</b> +9.2%
AE AudioCaps	64.2% - Qwen-A	<b>93.3%</b> +29.1%
CompA-R-test	80.0% - GAMA-IT	<b>96.4%</b> +16.4%
MuchoMusic	51.4% - Qwen-A	<b>56.5%</b> +5.1%
OpenAQA	80.0% - GAMA-IT	<b>86.0%</b> +6.0%
MusicInstruct (Long)	86.1% - MusiLingo	<b>90.2%</b> +4.1%
MusicQA	90.0% - MusiLingo	<b>93.0%</b> +3.0%
CMM Hallucination	76.0% - SALMONN	<b>82.0%</b> +6.0%
LongAudioBench ( <i>ours</i> )	45.3% - Gemini F v2	<b>64.2%</b> +18.9%

Results on expert reasoning tasks. AF2 outperforms all models by significant margins.

# AF2: Qualitative Results



**Question: What is unusual about the audio?**

**Ground Truth:** The audio contains combination of music and hammer beating which is unusual.

LTU: There is nothing unusual.

Qwen2-Audio-Instruct: The input audio contains a dark electronic piece with synthesizers and drums that has an energetic and experimental mood, instrumental only, and includes bass and percussion.

GAMA: The audio clip does not contain any unusual sounds. It is a typical example of a musical composition with drums and percussion instruments playing.

Gemini Pro v1.5: The unusual thing about the audio is that it sounds like someone is \*\*shredding paper by hand\*\*, but at a very \*\*fast and consistent pace\*\*. It's so fast and rhythmic that it almost sounds mechanical, but the slight variations and imperfections suggest a human hand. It's the speed and consistency that makes it stand out.

**Audio Flamingo 2 (*ours*):** the unusual aspect of the audio is the combination of the music and the sound of the mechanisms.

# AF2: Qualitative Results



**Question: When is the song with a male vocalist heard? Choose the correct option among the options below: (A) Towards the beginning (B) In the middle (C) Towards the end**

**Ground Truth: (B) In the middle**

LTU: Male vocalist can be heard throughout the audio.

Qwen2-Audio-Instruct: The male vocalist can be heard at the beginning of the track. Therefore, the answer is (A).

GAMA: (A) Beginning

Gemini Pro v1.5: The piano melody begins right at the start of the clip. So the answer is (A).

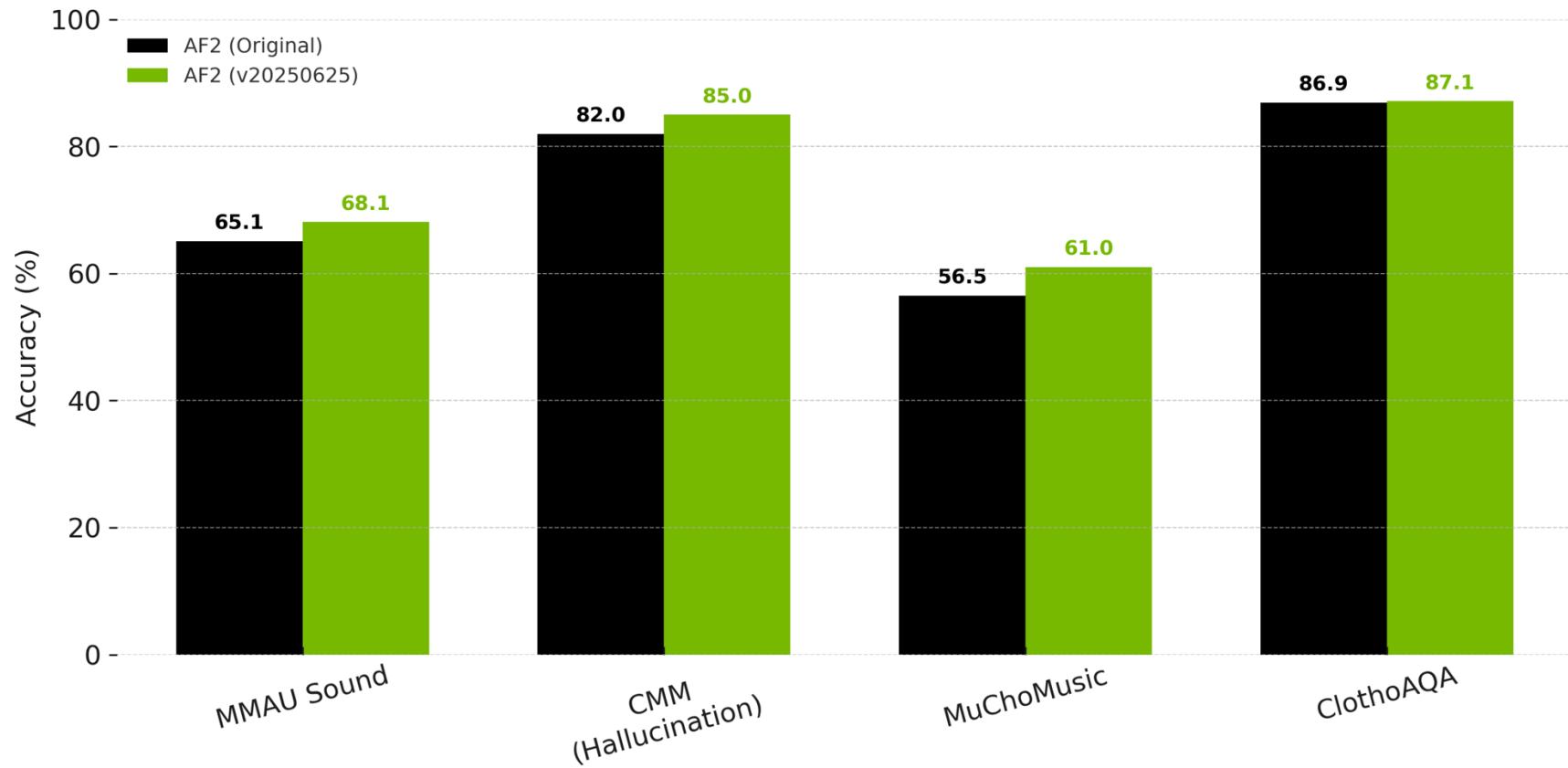
Audio Flamingo (chat): End

**Audio Flamingo 2 (*ours*): (b) middle**

# AF2: New Checkpoints

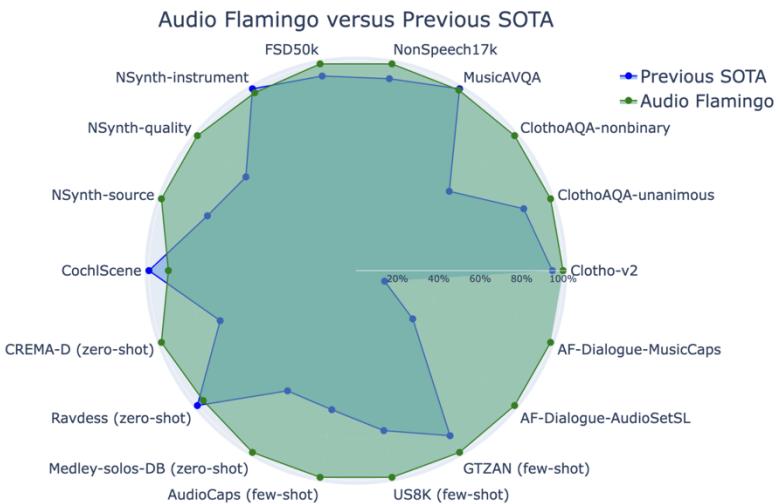


We just released a better and stronger AF2 (v20250625)! Try it out!

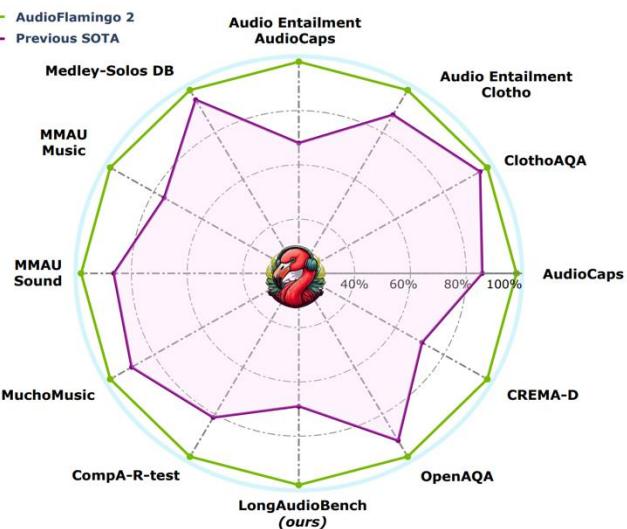


# Audio Flamingo Series

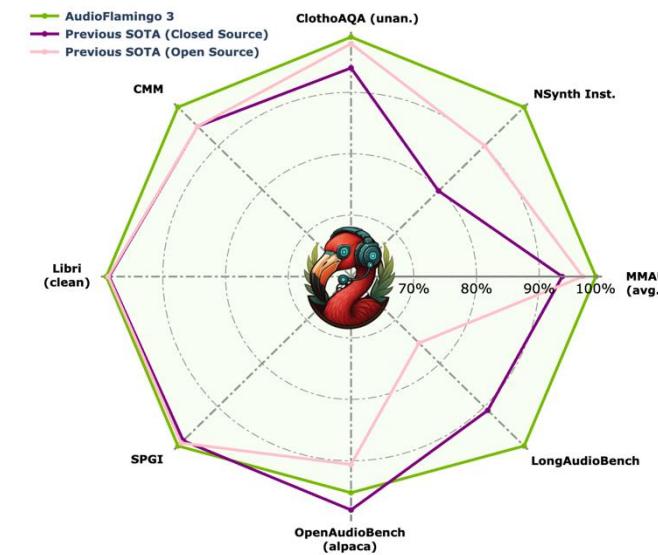
## Audio Flamingo 1 (ICML 2024)



## Audio Flamingo 2 (ICML 2025)



## Audio Flamingo 3 (Under review)



- Focus on foundational audio processing tasks (e.g., classification, captioning, QA, etc.)
- Introduced **RAG** and **multi-turn dialogue** abilities
- ~**5.9 M** training instances

- Focused on improving performance on **expert reasoning benchmarks**
- Introduced **long-audio understanding and reasoning**
- ~**10 M** training instances and synthetic data

- Focused on improving performance **across all tasks by scaling up on open-source data**
- Introduces **speech** in the Audio Flamingo series
- Introduces **long-speech understanding and reasoning, thinking abilities and multi-turn, multi-audio chat**
- ~**50 M** training instances

# Training Data

A Mix of Highly Heterogenous Open-sourced Datasets and Synthesized Datasets

## Existing Datasets:

### Speech:

Datasets: 10  
QA Pairs: ~6M  
Hours: ~18.4K

### Sound:

Datasets: 12  
QA Pairs: ~10M  
Hours: ~10K

### Music:

Datasets: 10+  
QA Pairs: ~400K  
Hours: ~30K

### AudioSkills-XL (Ours):

Datasets: 11  
QA Pairs: 12M+  
Hours: 50K+

### LongAudio-XL (Ours):

Datasets: 10+  
QA Pairs: ~1.25M  
Hours: 5K+

### AF-Think (Ours):

Datasets: 5  
QA Pairs: 350K+

### AF-Chat (Ours):

Datasets: 3  
Instances: 50K+



## 1. Foundation model

audio (<=30 seconds) + instruction → output

**Input:**

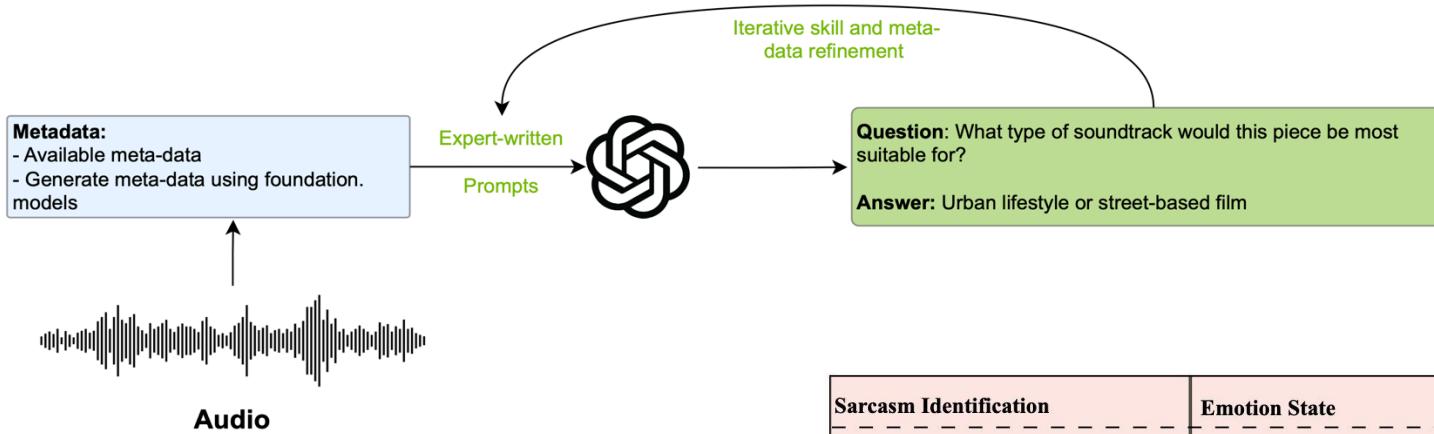
Describe this audio in a sentence.

**Audio Flamingo 3:**

a man is speaking and then a loud bang occurs.

# AudioSkills

The 1st attempt at creating a large-scale Audio-QA dataset, focused on expert-skills



Other AQA Datasets
<b>Foundational:</b> What is the genre of the music? <b>Answer:</b> Rock
<b>ClothoAQA:</b> Are there people having conversation? <b>Answer:</b> Yes
<b>OpenAQA:</b> What kind of sound do the mechanisms make? <b>Answer:</b> Mechanical sounds
<b>SALMONN:</b> What is happening to the vehicle? <b>Answer:</b> Vehicle is accelerating.

Temporal Relationship Identification	General Reasoning	Counting
<b>Order:</b> In what sequence do the sounds first appear in the audio? (A) Rain (B) Human voice ....	<b>Question:</b> How does the melody in the audio contribute to the hypnotic effect of the music? (A) By changing frequently to maintain interest. (B) By maintaining a consistent and repetitive loop ....	<b>Level 1 &amp; 2:</b> How many times did the dog bark? <b>Answer:</b> Two
<b>Attribute:</b> How does the sound of thunder change over time? (A) Gets louder (B) Gets softer ....	<b>Attribute Identification</b>	<b>Level 3:</b> How many times did the second sound occur in the entire audio? <b>Answer:</b> Four
<b>Grounding:</b> When does the human voice appear in the audio? (A) Beginning (B) Middle (C) End		
<b>Referring:</b> What sound appears last in the audio? (A) Rain (B) Human voice ....		

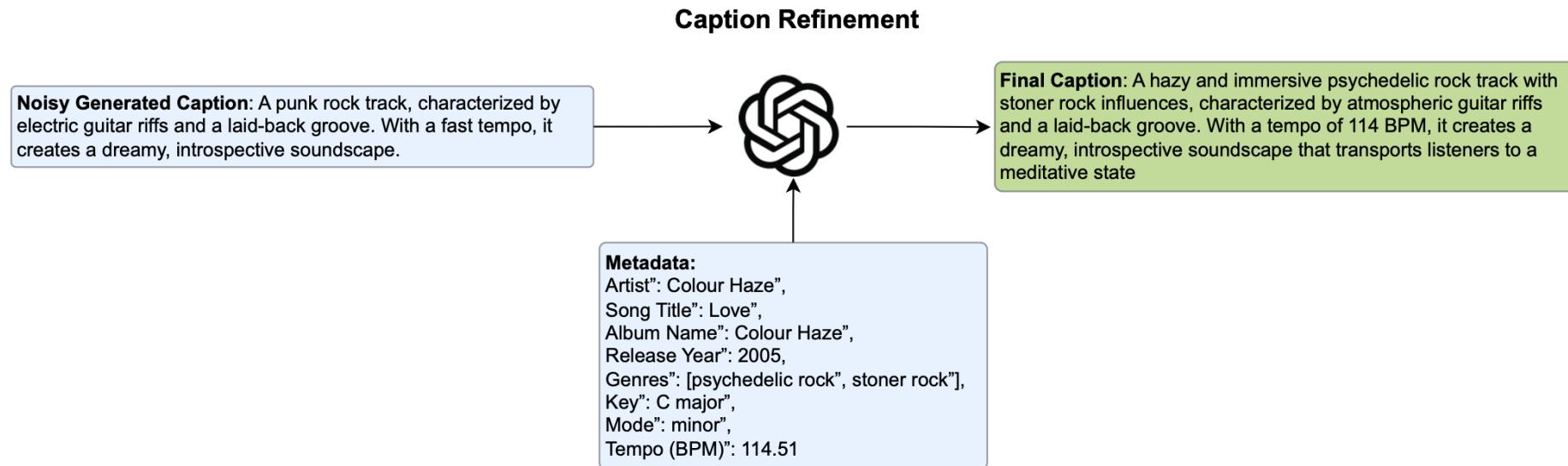
Contextual Speech Event Reasoning	Contextual Sound Event Reasoning
<b>Question:</b> How does the repetition in the spoken utterance influence the interpretation of the sequence of actions taking place? <b>Answer:</b> The repetition suggests urgency or frustration in an effort to manage an ongoing situation, possibly highlighted by background sounds of movement.	<b>Question:</b> Considering the correlation of the man's speech and the music, infer his possible connection to the music playing in the scene. <b>Answer:</b> The man is likely a guitarist or an instructor who is explaining the process of tuning a guitar.

Sarcasm Identification	Emotion State	Information Extraction
<b>Question:</b> In the conversation, why might the suggestion 'You're smart. You could go back to school--finish your Master's that you were started never finished' be considered sarcastic?	<b>Identification:</b> In the input conversation, when discussing a forgotten item, what was the speaker's emotional state while acknowledging their oversight?	<b>Needle QA:</b> What was the specific example given to highlight the difference between investing and speculating with cryptocurrencies?
<b>Topic Relationship Rasoning</b>	<b>Causal Reasoning:</b> In the input conversation, why does the speaker feel excited while describing the fish phenomenon?	<b>Causal QA:</b> What caused one speaker to become frustrated about the bag situation?
<b>Information Extraction</b>	<b>Response QA:</b> How does the speaker respond when asked about their study program?	
Order	Temporal Order	Temporal Attribute
	<b>Temporal Order:</b> What is the order in which the speaker discusses the topics in the speech? (A) The speaker introduces the last crime and its seasonal timing. ....	<b>Temporal Attribute:</b> How does the focus of the speech evolve over time? (A) It moves from general grievances to more specific legislative conflicts., (B) It starts optimistic and becomes critical. ....
	<b>Temporal Referring:</b> When does the speaker discuss the revival of mills and factories in relation to other topics? (A) At the very beginning, (B) After describing the general social mood, ...	<b>Temporal Grounding:</b> At what point in the speech does the speaker describe the specific setting of the last crime? \nChoose the correct option from the following options: (A) At the beginning, (B) In the middle, (C) At the end

## Synthetic Knowledge Generation for Audio LALMs

- Music is a knowledge-intensive domain, unlike speech and sounds that can be perceived and understood by most humans.
- Music datasets are the most low-resource across all 3 domains.

Solution: Transfer world knowledge from text to improve music understanding.



## 2. Long audio understanding

Audio (up to 10 mins) + instruction → output

**Input:**

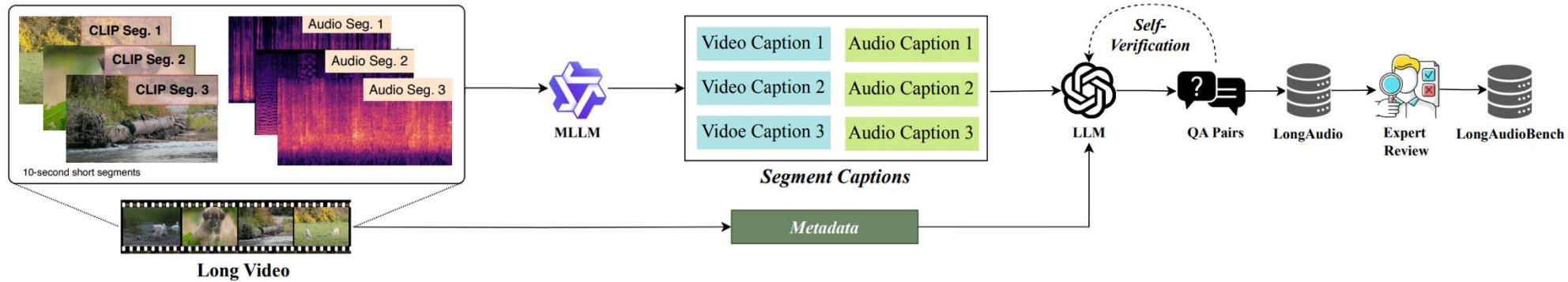
🔊 How does the tone of female speech change throughout the audio?

**Audio Flamingo 3:**

Sad to Happy

# LongAudio

The 1st attempt at creating a large-scale long Audio-QA dataset



~1.25M QA samples, 10+ skills, 30 seconds - 10 minutes, sourced from 10+ datasets

## Long-Form Sound and Music QA

**Temporal Question:** What can be heard at the end of the audio? Choose the correct option among the options below:  
(A) A man speaking in English, (B) Classical music with drums, (C) Instrumental electronic piece, (D) A pop song  
**Answer:** (B) Classical music with drums

**Sub-scene Question:** What do you hear between the piano playing and the guitar strumming?  
**Answer:** A male voice speaking in English with a sad mood.

## Long-Form Speech QA

**Sarcasm Identification Question:** In the given spoken conversation, is the comment about "it'll be better that way" made by a speaker sarcastic?  
**Answer:** No, the comment is not sarcastic. The speaker is trying to comfort the other person by suggesting that time will help heal the pain.

**Information Extraction Question:** What is the response to the request for transferring the form information?  
**Answer:** The response is that it cannot be transferred, and everyone else is doing the same thing.

### 3. Audio reasoning (thinking)

Audio (up to 10 mins) + instruction → output

**Input:**



Please think and reason about the input audio before responding.

**Audio Flamingo 3:**

The audio contains sounds of a ball being hit and bounced, along with the sound of shoes squeaking on a hard surface. These sounds are characteristic of a sport played on a hard court with a racket or shuttlecock. Tennis and badminton are both such sports. The presence of a ball being hit and bounced suggests a sport where a racket is used to hit a ball, which is more indicative of tennis than badminton.

Therefore, the most likely sport is tennis



Which genre best describes the sound texture of this music piece? Choose one among the following options:

- (A) Pop rock
- (B) Deathcore
- (C) Classical symphony
- (D) Jazz fusion

Please think and reason about the input music before you respond.

Based on the intense, distorted guitar riffs, rapid-fire drumming (including what sounds like blast beats), and extremely low, guttural vocals, the music fits the characteristics of extreme metal subgenres. Comparing these sounds to the options provided, **\*\*(B) Deathcore\*\*** is the most accurate description.



## 4. AF3-Chat

multi-audio, multi-turn chat input → output text or voice

**Input 1:**

·||... Describe this audio in a sentence.

**Audio Flamingo 3:**

A man is speaking and then a loud bang occurs.

**Input 2:**

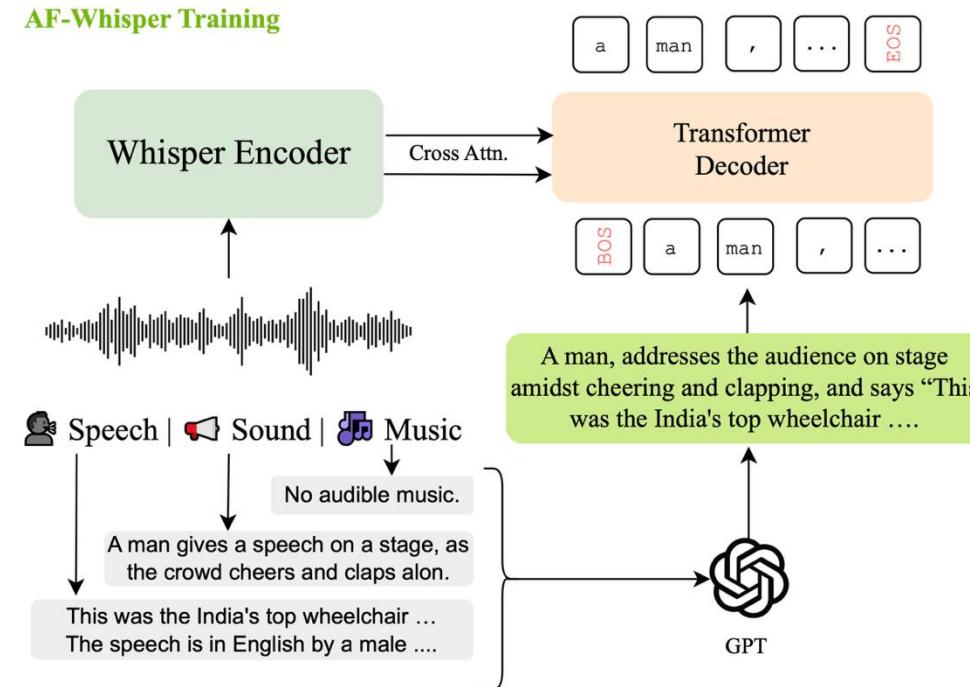
·||...

**Audio Flamingo 3:**

The bang is louder and there is no one speaking. ·||...

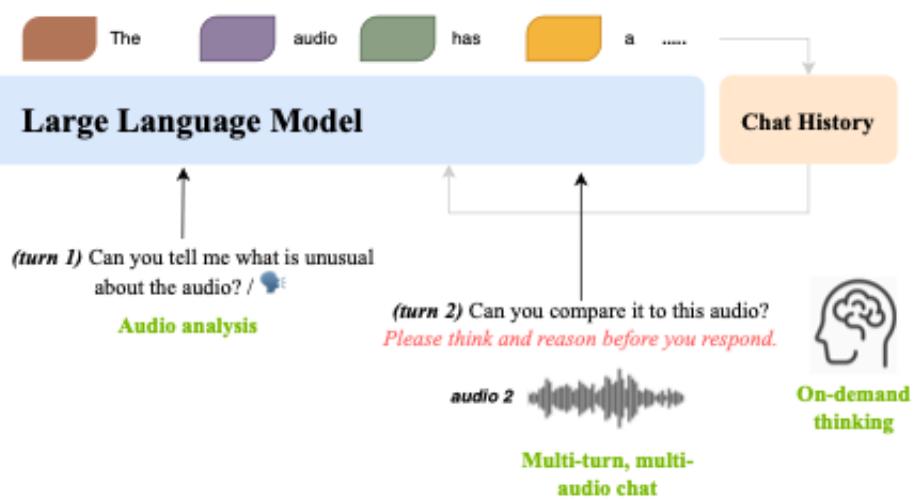
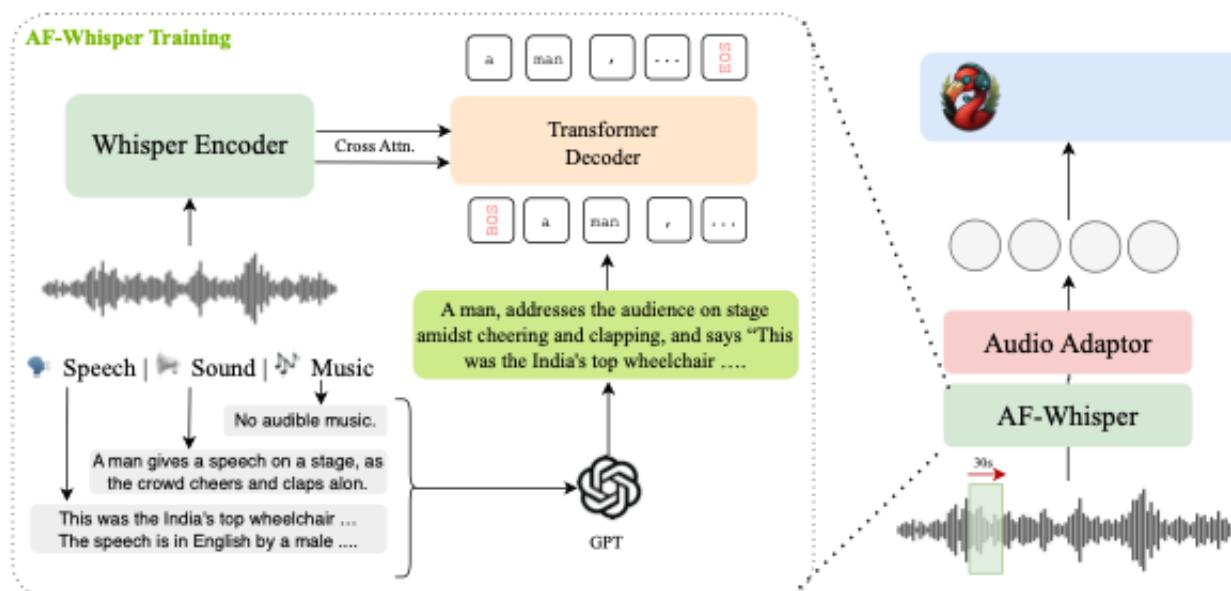
# Model Architecture

## AF-Whisper Training Strategy



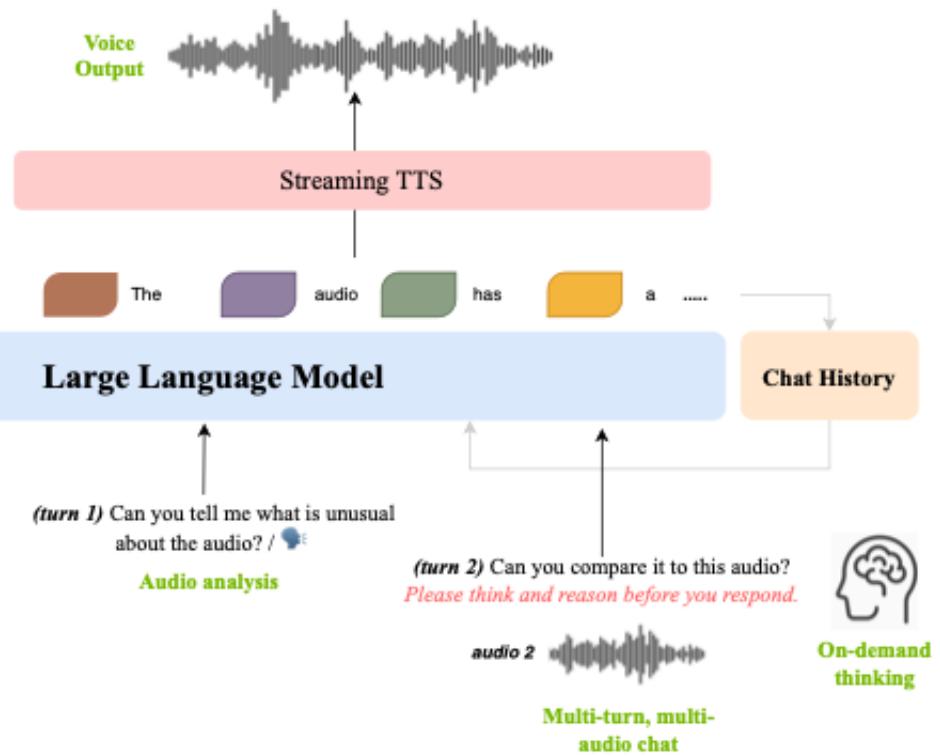
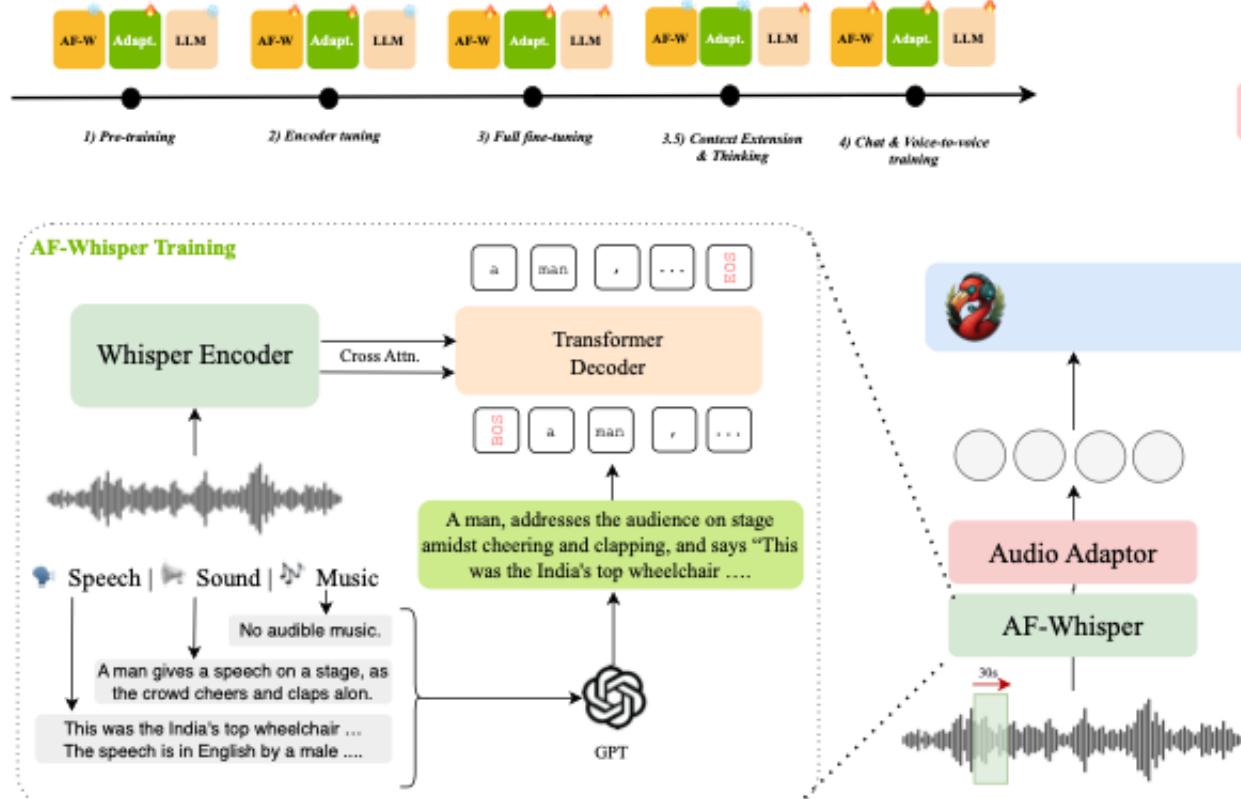
# Model Architecture

## Audio Flamingo 3



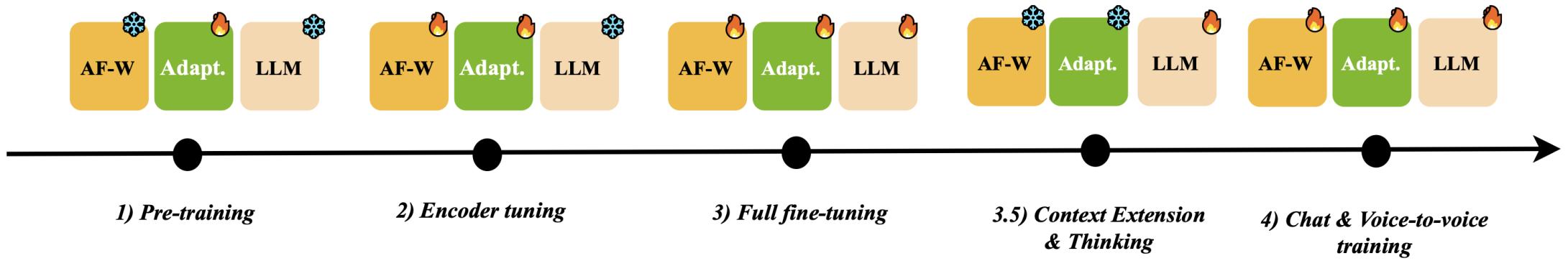
# Model Architecture

## Audio Flamingo 3 with Voice



# Training Stages

Audio Flamingo 3 has 5 training stages



# Results

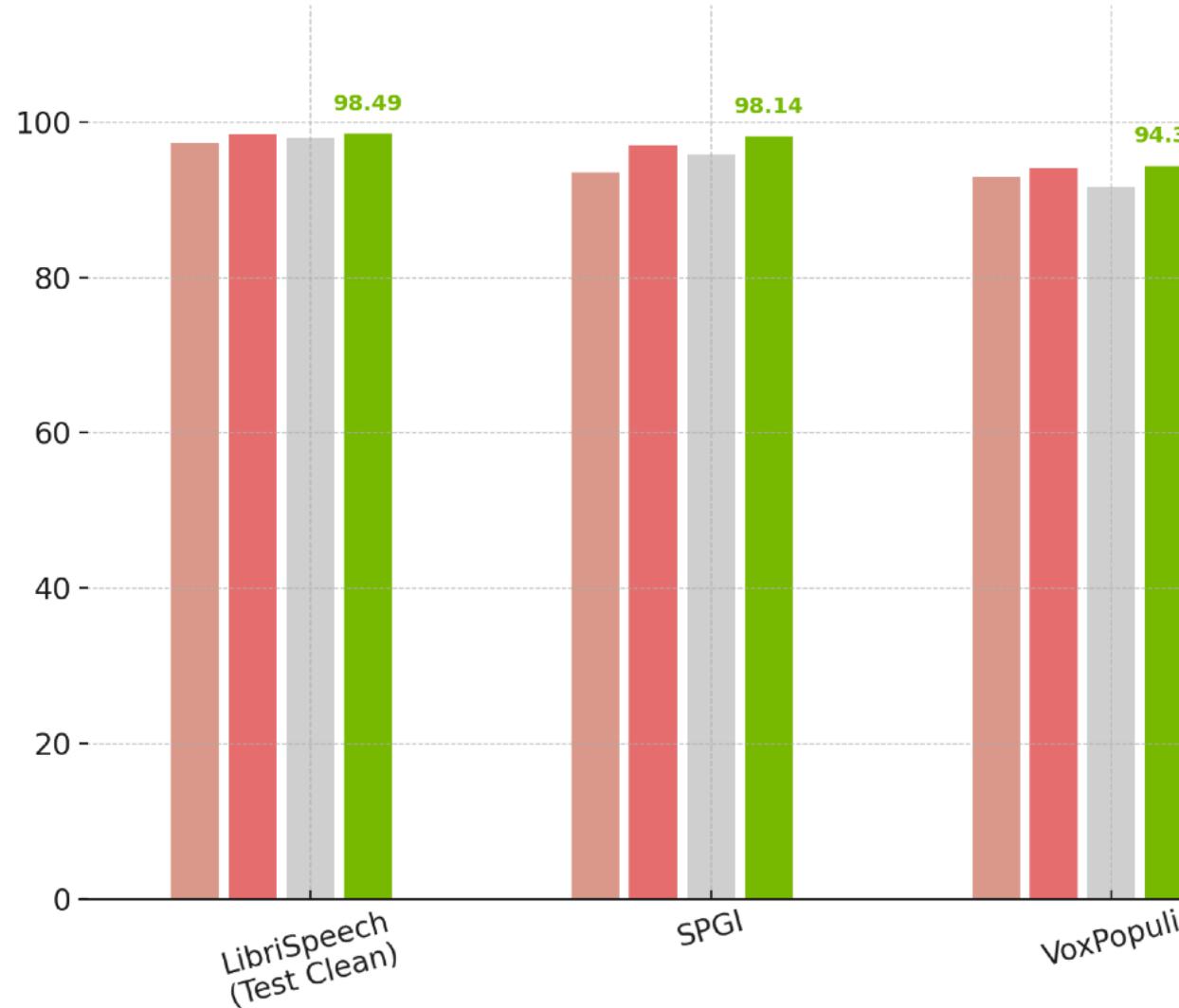
AF 3 achieves competitive results compared to open-source / closed source models in ASR



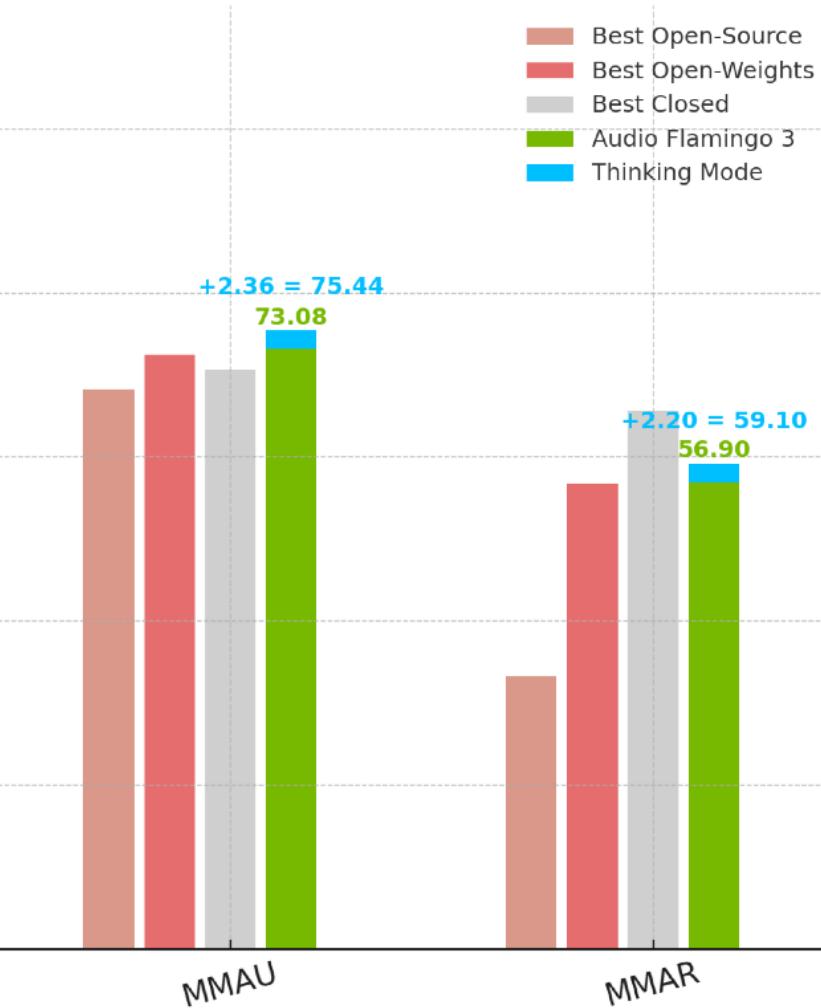
# Results

AF3 achieves SOTA in expert audio understanding and reasoning tasks

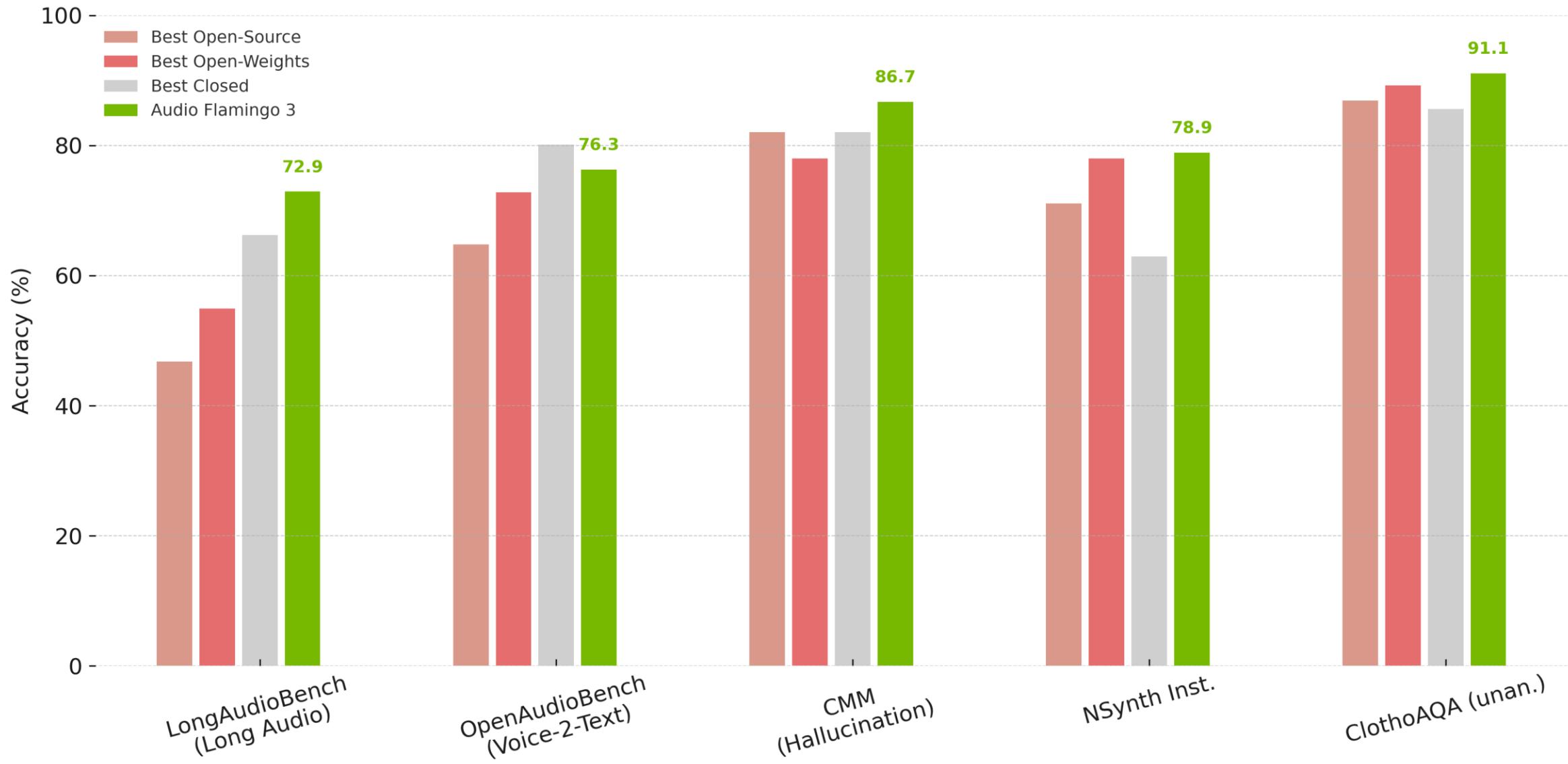
Automatic Speech Recognition (100-WER)



Expert Reasoning (Accuracy)

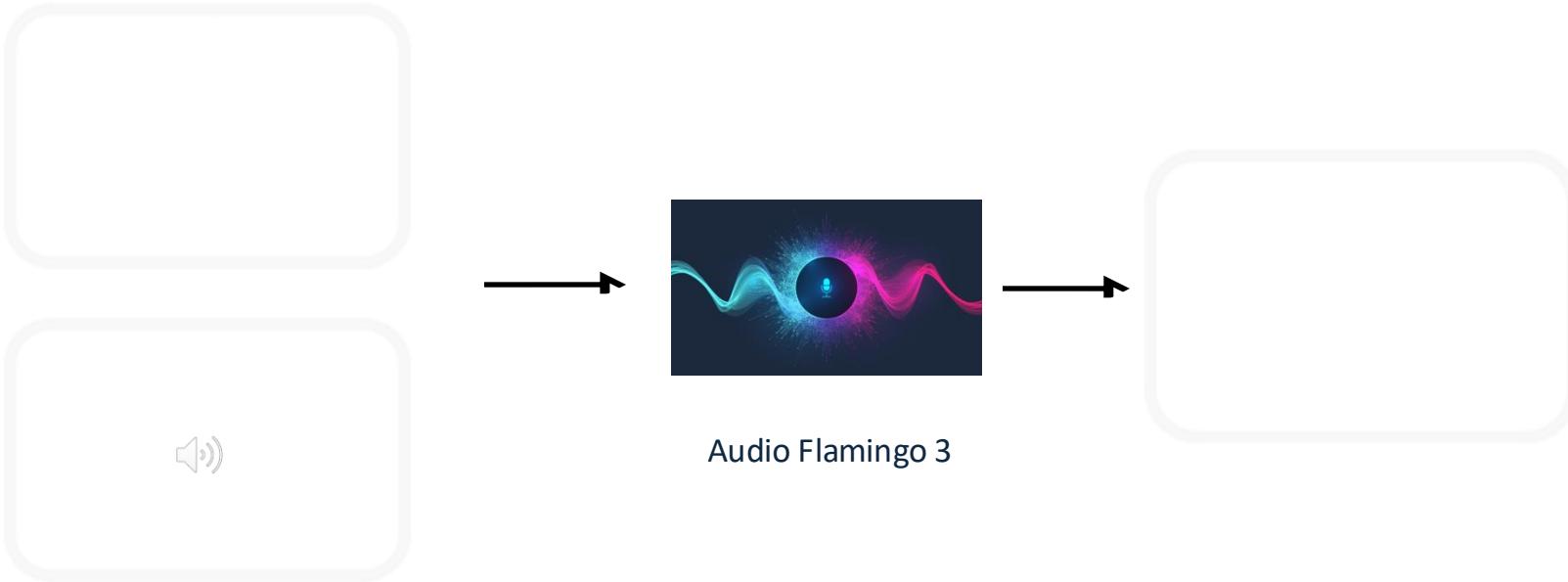


# Results



# Demos

AF3 has Emergent Capabilities



# Demos

AF3 has Emergent Capabilities

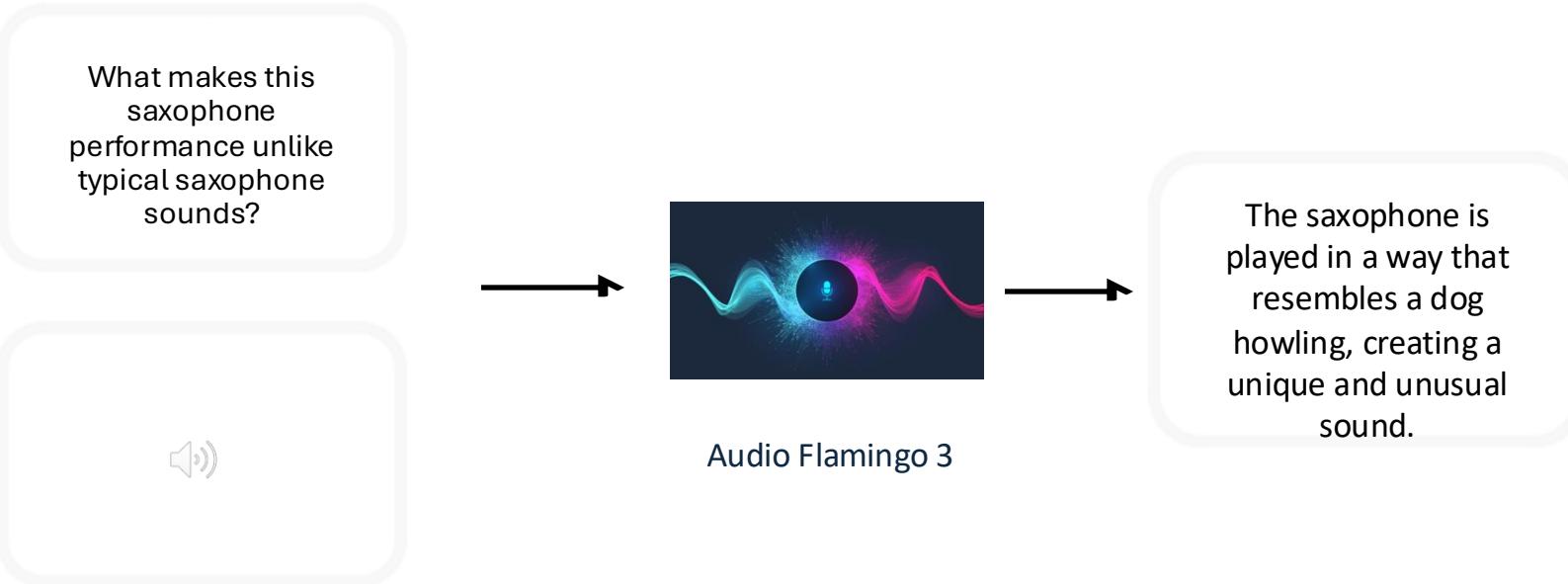
What makes this  
saxophone  
performance unlike  
typical saxophone  
sounds?



Audio Flamingo 3

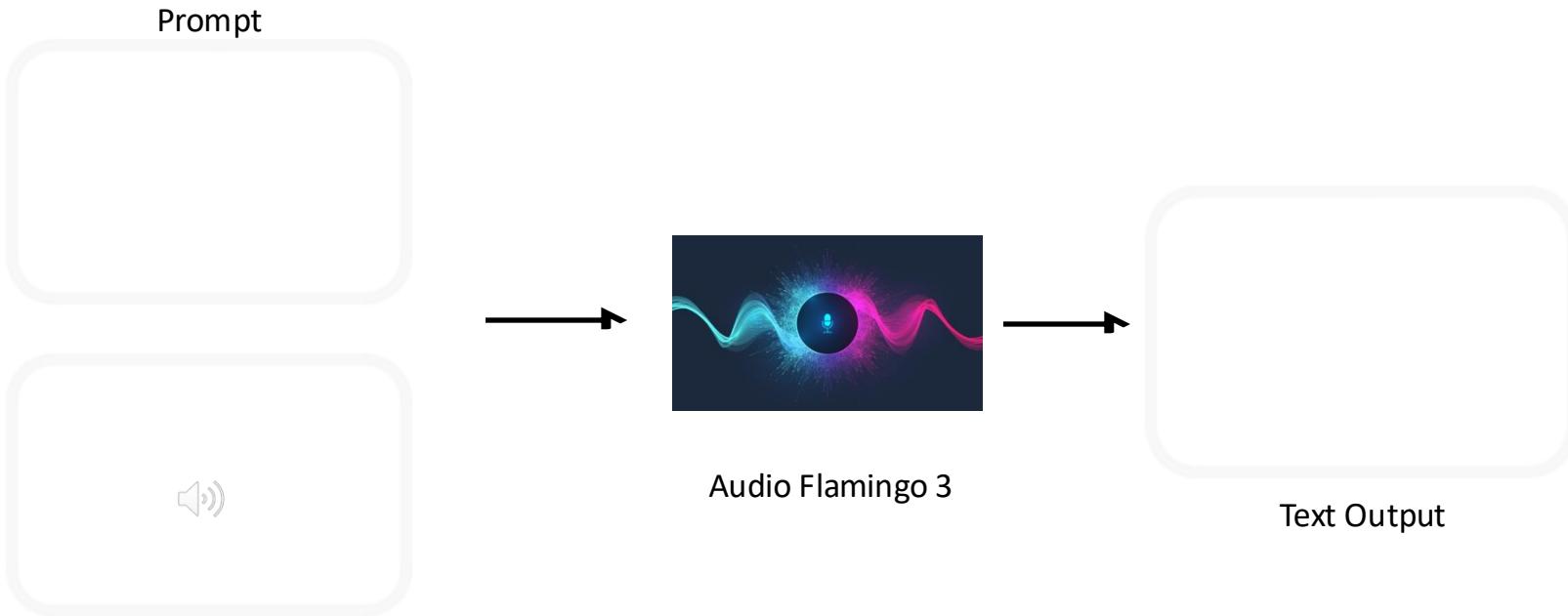
# Demos

## AF3 has Emergent Capabilities



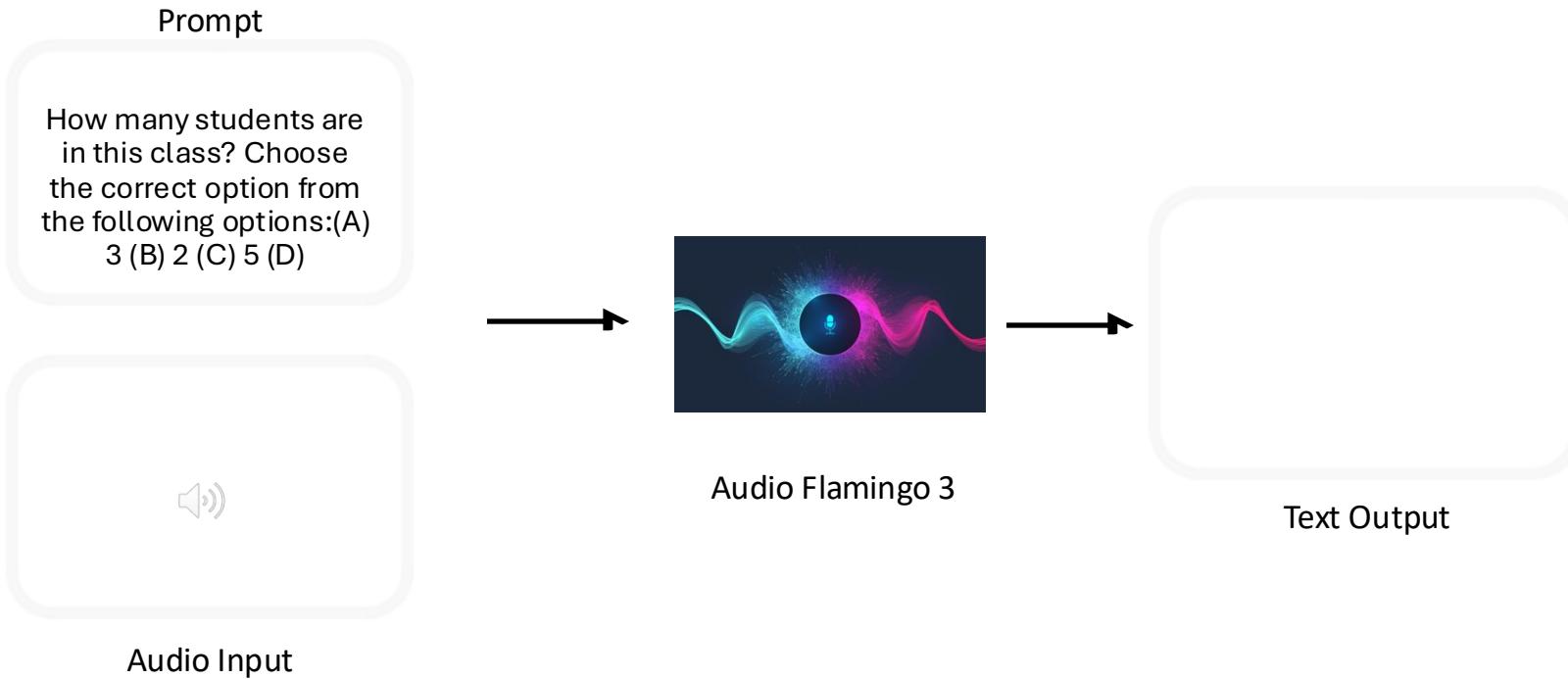
# Demos

AF3 has thinking and reasoning capabilities



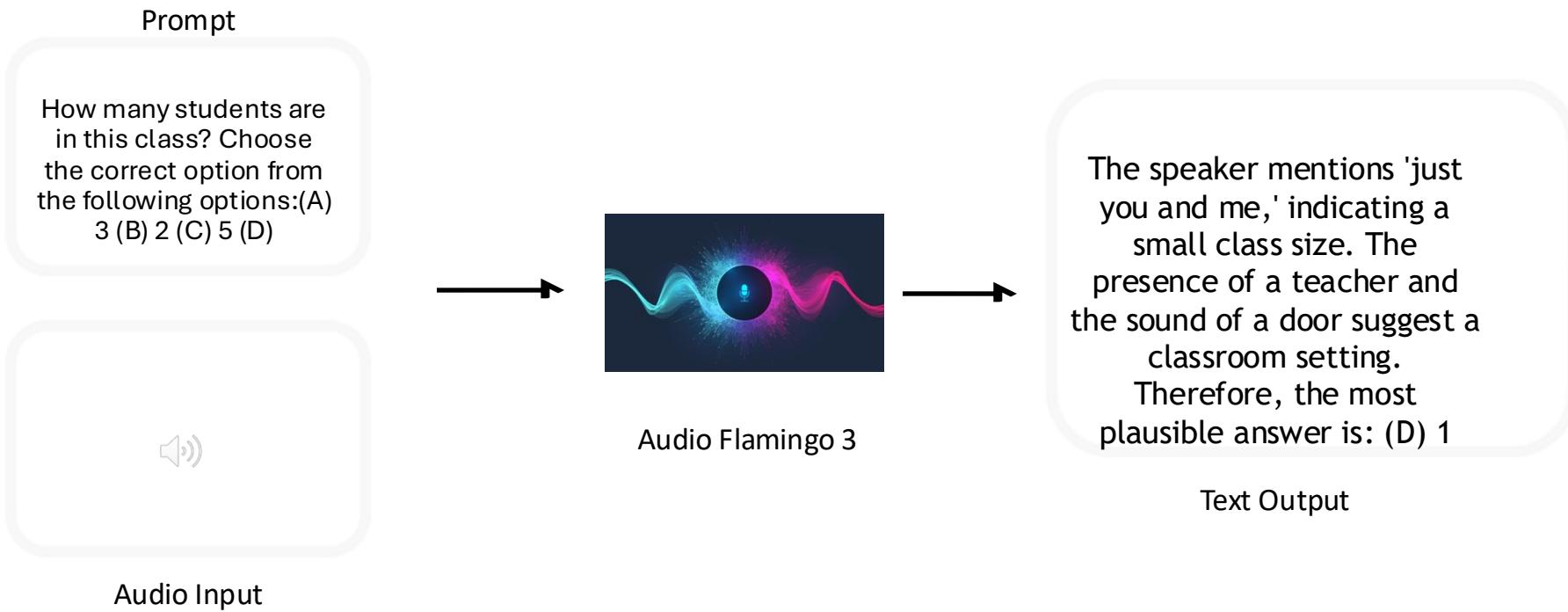
# Demos

AF3 has thinking and reasoning capabilities



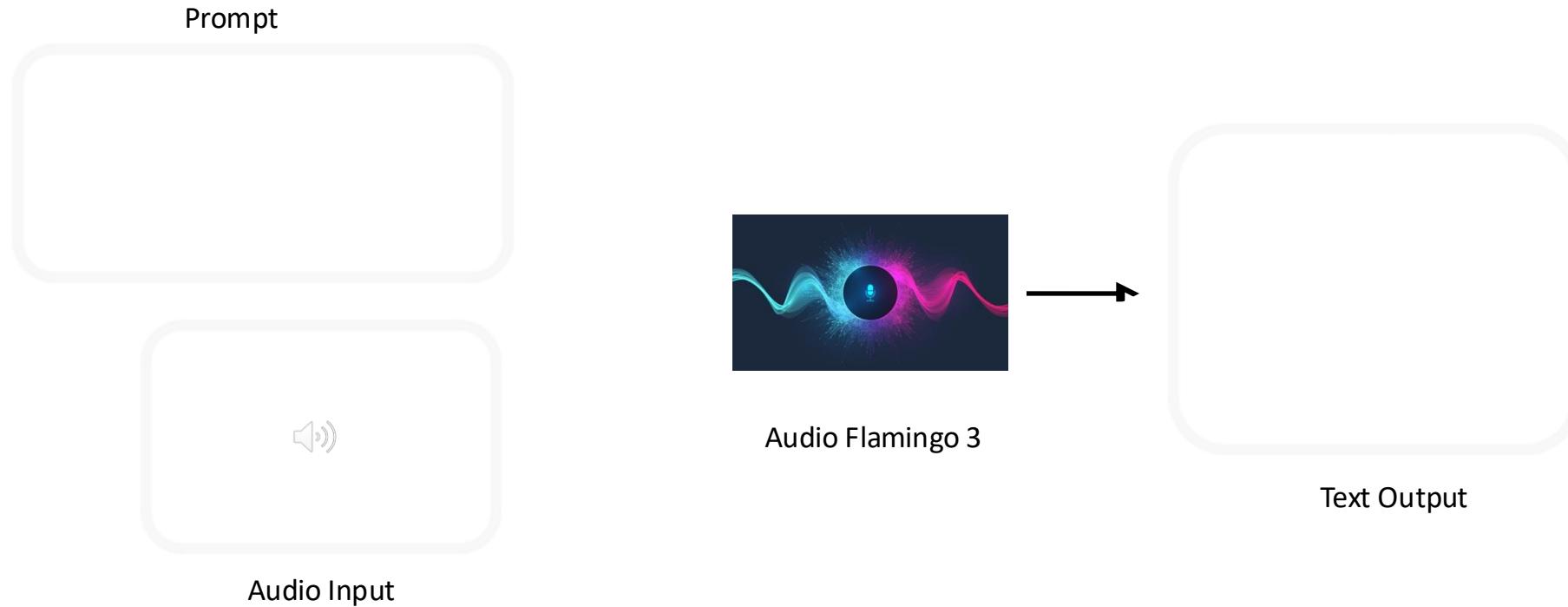
# Demos

AF3 has thinking and reasoning capabilities



# Demos

AF3 combines speech and sound captions into one



# Demos

AF3 combines speech and sound captions into one

Prompt

Generate a caption describing the auditory scene in the input audio, considering spoken content, background sounds, and music.



Audio Input

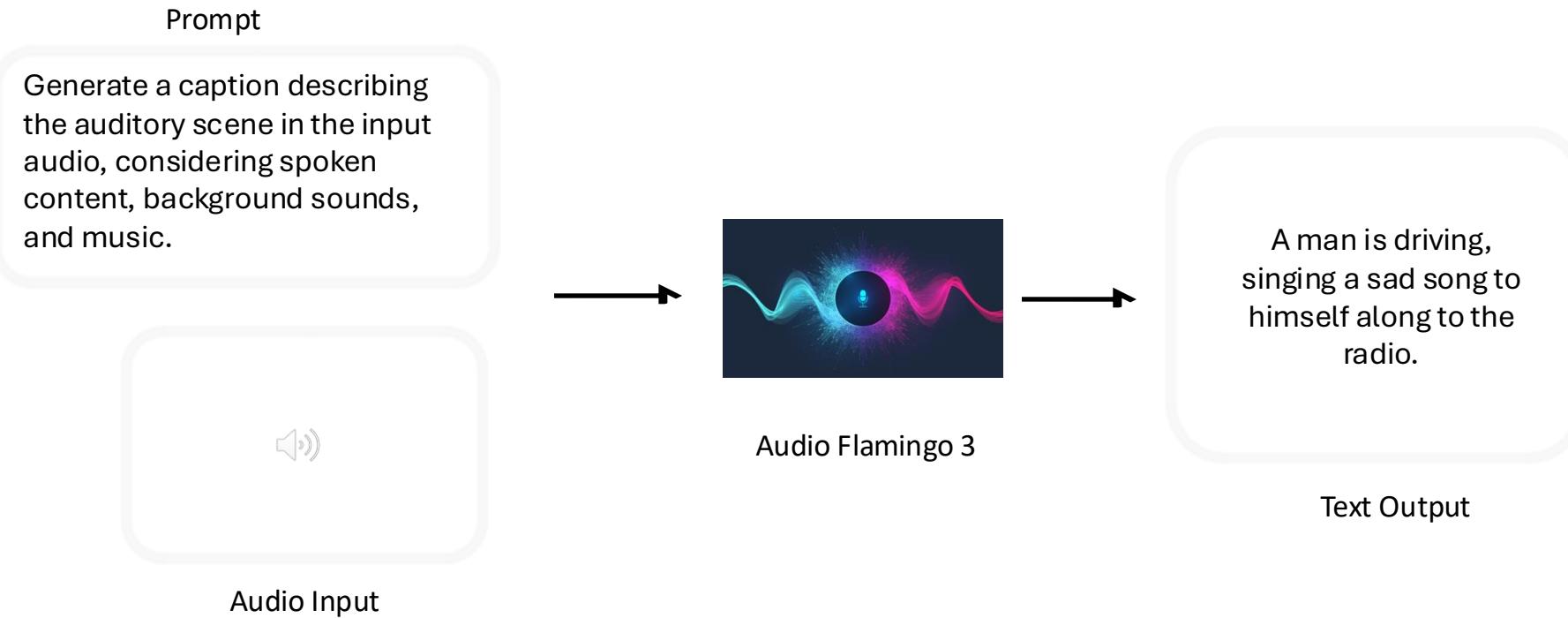


Audio Flamingo 3

Text Output

# Demos

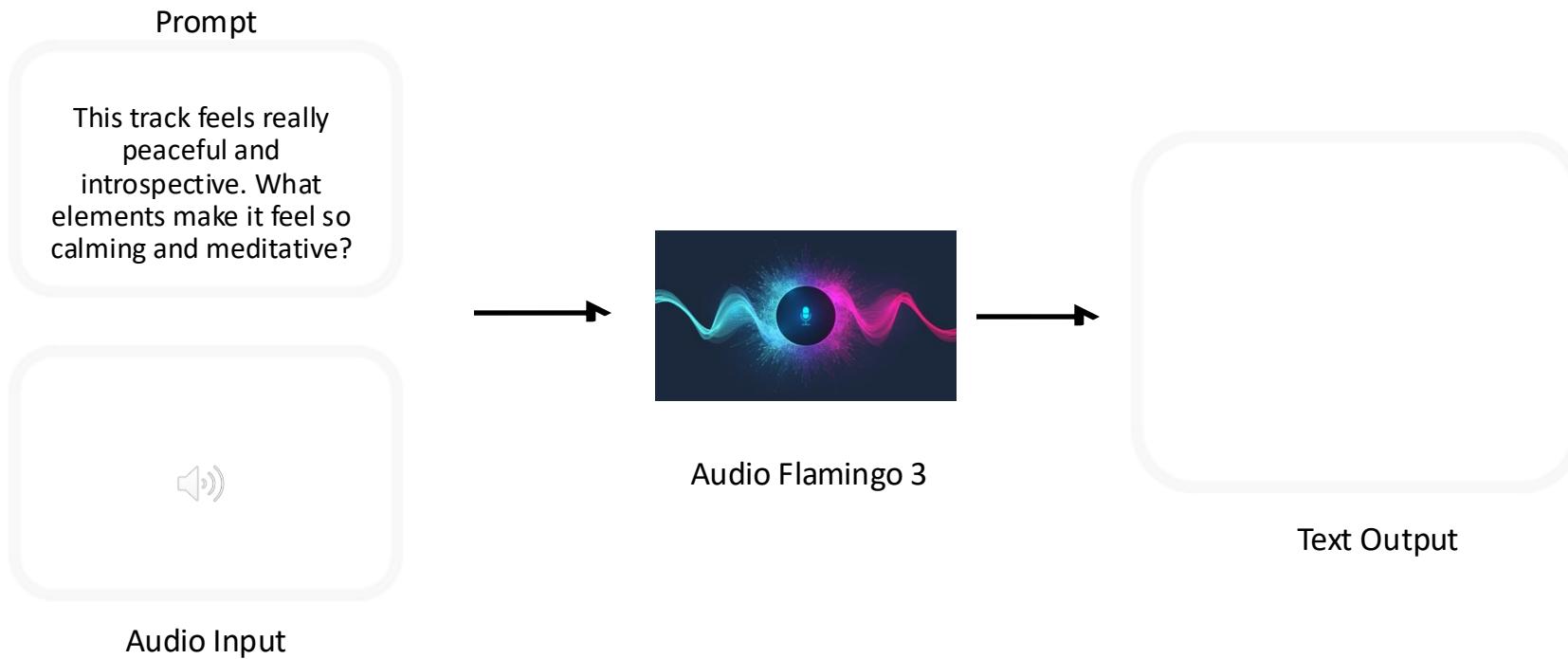
AF3 combines speech and sound captions into one



ASR only transcription: “Nothing really matters”  
Audio only caption: Sound of a car and music in the background

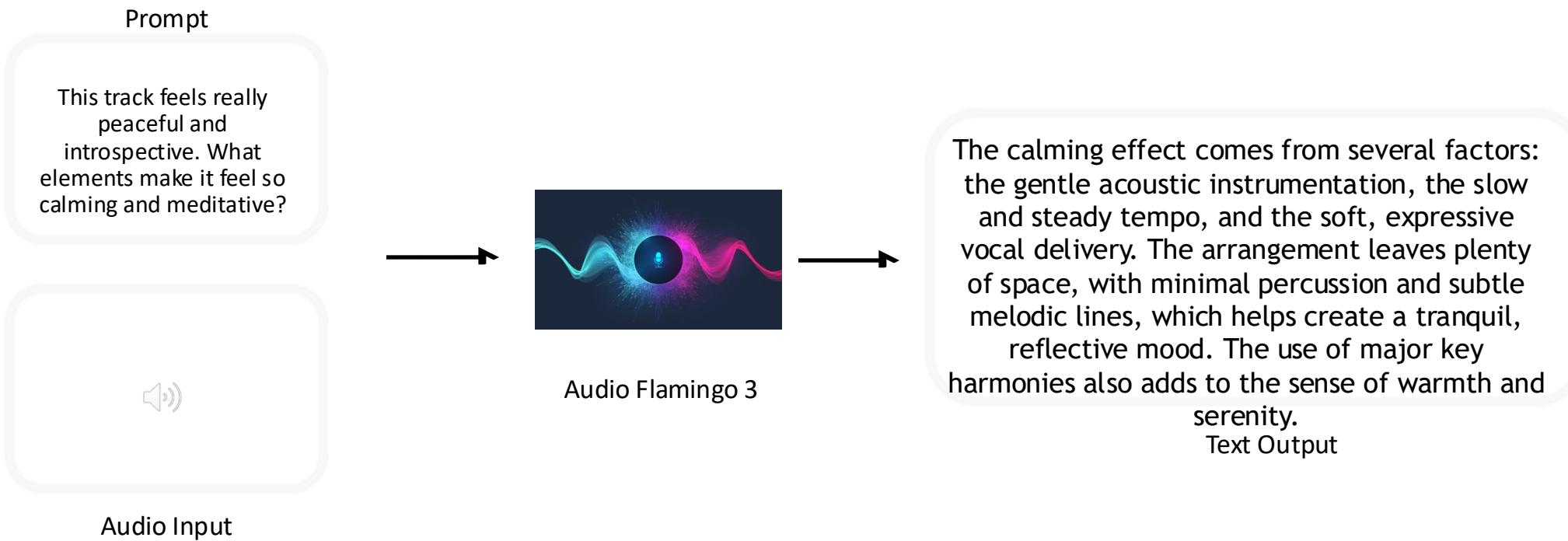
# Demos

With AF3, users can chat with multiple audios



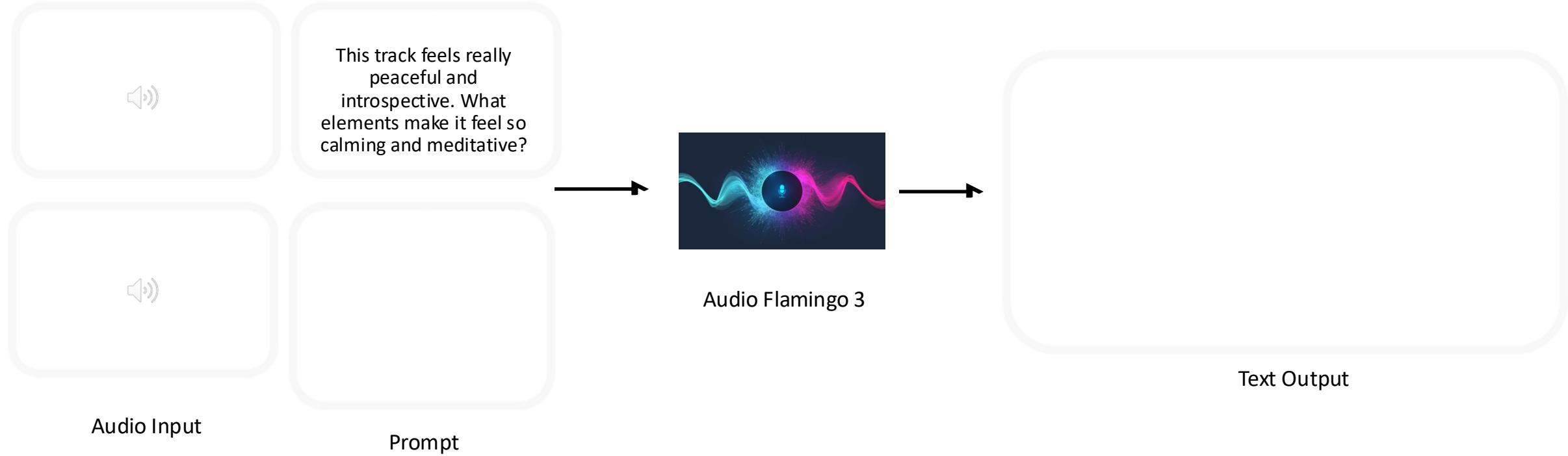
# Demos

With AF3, users can chat with multiple audios



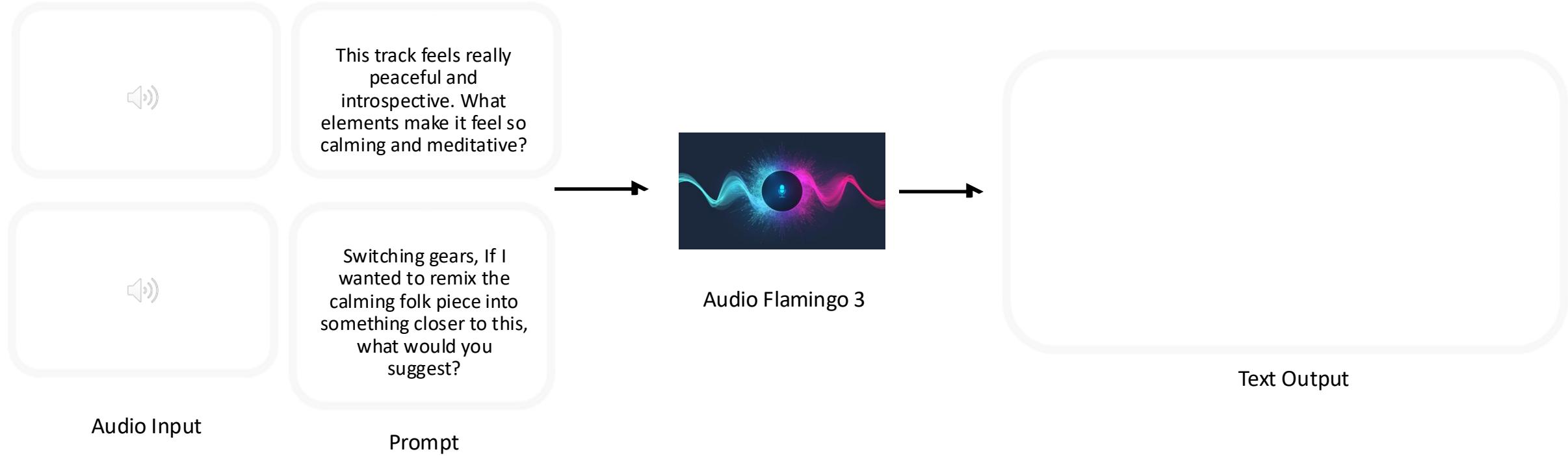
# Demos

With AF3, users can chat with multiple audio inputs



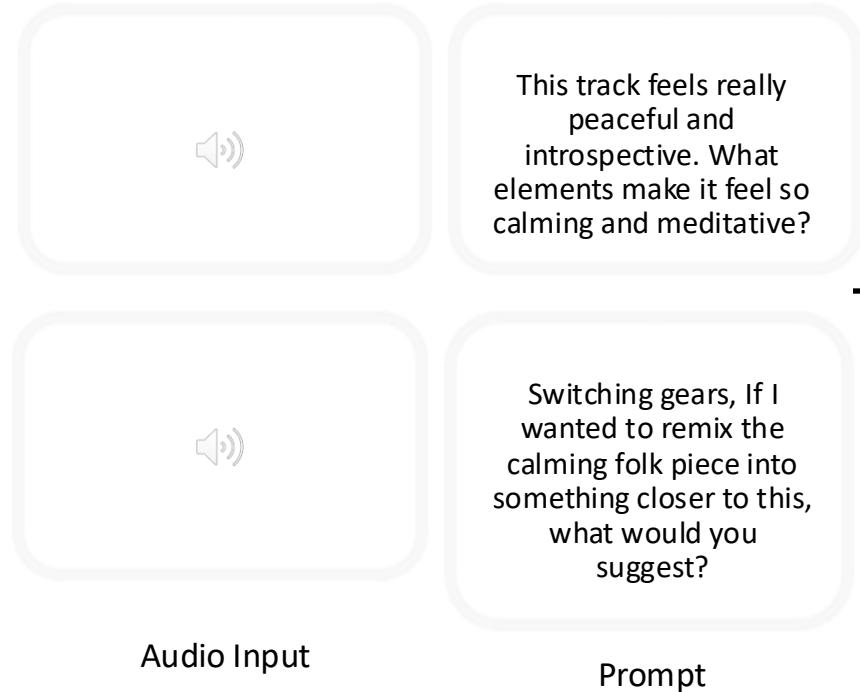
# Demos

With AF3, users can chat with multiple audio inputs



# Demos

With AF3, users can chat with multiple audio inputs



Audio Flamingo 3

To transform the folk track into a more energetic, synth-driven style, start by increasing the tempo and introducing electronic drums with a punchy, danceable groove. Replace or layer the acoustic instruments with bright, rhythmic synths and arpeggiators. You could also process the vocals with effects like vocoder or light autotune for a more electronic feel. The key is to maintain the melodic core but reframe it with bold, synthetic textures and a driving beat.

Text Output

# Demos

AF3 has voice-to-voice interaction



# Demos

AF3 has voice-to-voice interaction



<https://audioflamingo3.github.io/>

# Ongoing Future Work

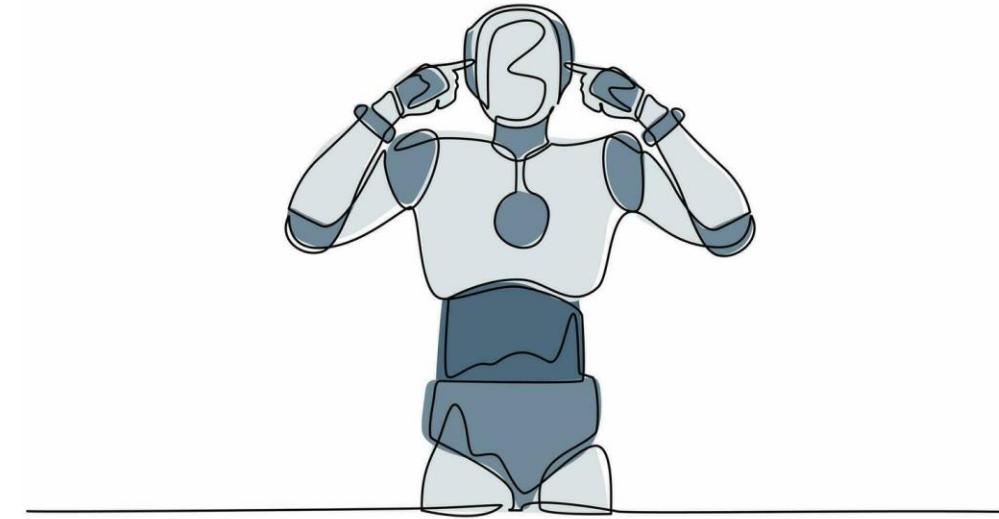
- **Combining audio generation and understanding**
  - Research Question 1: Can audio generation and understanding help each other?
  - Research Question 2: How to effectively build models that can both understand and input audio and generate one?
- **Thinking Models:**
  - Research Question: Scaling inference-time compute for audio generation and understanding by moving to the thinking paradigm
- **Making Audio-Language Models more efficient:**
  - Research Question: How far can we go with small language models?
- **Audio-visual complex reasoning:**
  - Research Question 1: How do we generate instruction-tuning datasets for complex reasoning over audio-visual cues?
  - Research Question 2: Leveraging findings from our audio work, how do we build effective architectures for effective audio-visual reasoning?
- **Music understanding and generation**

# Audio Understanding on Physical Devices

## Smartglasses



## Robots with Hearing Capabilities



# Thanks to all our sponsors and collaborators!



**Adobe**



**Microsoft**

**Dolby**



**NVIDIA**®



**Thank You!**

**Questions?**