

Faithful and Grounded Audio Language Models

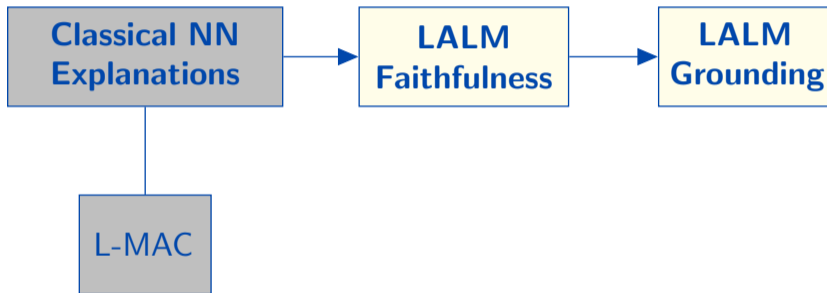
Cem Subakan

Conversational AI Reading Group, 2026

April 9, 2026



Plan



The last time vs this time

Posthoc Explanations for Audio Models

Cem Subakan

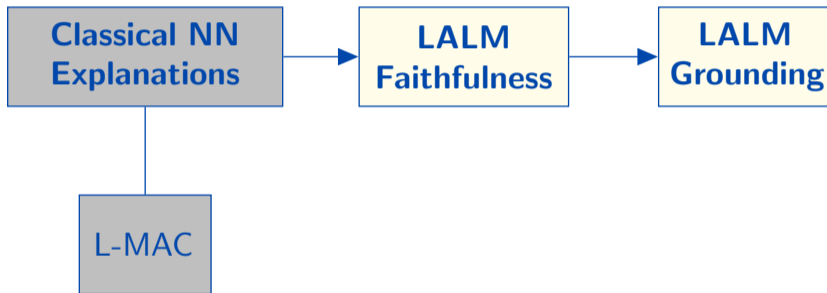
December 5, 2024



0:03 / 1:01:28

1 / 34

Plan



Students leading

Pooneh Mousavi



Lovenya Jain



Jihoon Jeong (Huni) Francesco Paissan



Explainable Machine Learning

- Black-box models

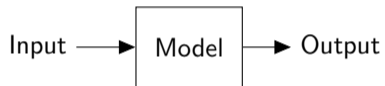


Explainable Machine Learning

- Black-box models



- Explainable Models

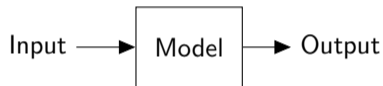


Explainable Machine Learning

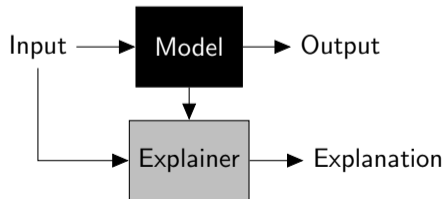
- Black-box models



- Explainable Models



- Posthoc Explanations

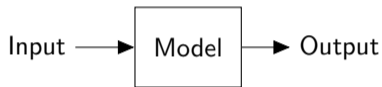


Explainable Machine Learning

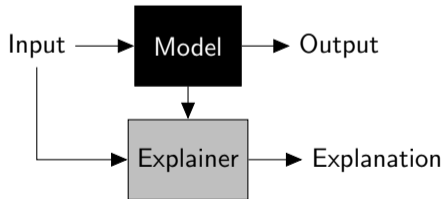
■ Black-box models



■ Explainable Models



■ Posthoc Explanations



Desiderata: Faithful, Listenable, Understandable Explanations

Important Tool for Decision Critical Applications (e.g. Healthcare, DeepFake detection)

Table of Contents

Neural Network Explanations

- Listenable Maps for Audio Classifiers
- Faithfulness

Faithfulness in LALMs

- Adversarial Interventions
- Noise and Masking Interventions
- Guided Masking
- CoT Interventions

Grounding the LALM Reasoning

Table of Contents

Neural Network Explanations

- Listenable Maps for Audio Classifiers

- Faithfulness

Faithfulness in LALMs

- Adversarial Interventions

- Noise and Masking Interventions

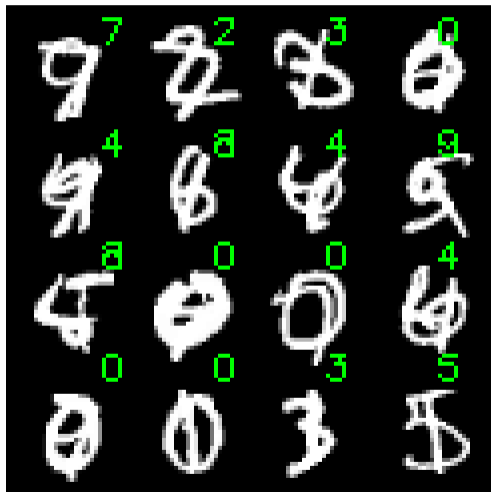
- Guided Masking

- CoT Interventions

Grounding the LALM Reasoning

Neural Network Explanation

- Why does this particular input lead to that particular output?



Neural Network Explanation

- Why does this particular input lead to that particular output?



Explanations

- Saliency maps are commonly used in computer vision for producing explanations (e.g. GradCAM, Integrated Gradients, GradSHAP, SmoothGrad, ...).



Explanations

- Saliency maps are commonly used in computer vision for producing explanations (e.g. GradCAM, Integrated Gradients, GradSHAP, SmoothGrad, ...).

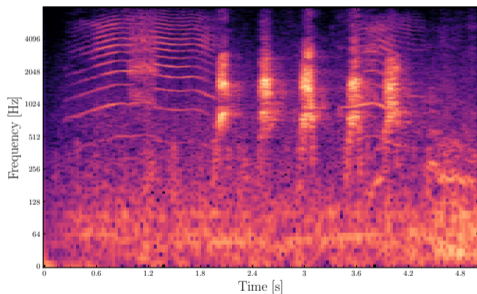


- The explanations should **faithfully** follow the original model.

Audio domain NN explanations

- We developed a **listenable** and **faithful** (SOTA) posthoc explanation method for audio classifiers.

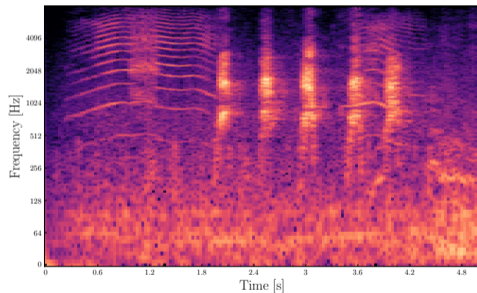
Input Audio



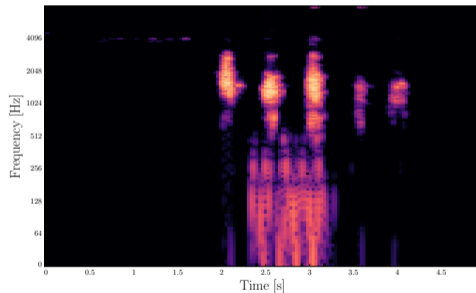
Audio domain NN explanations

- We developed a **listenable** and **faithful** (SOTA) posthoc explanation method for audio classifiers.

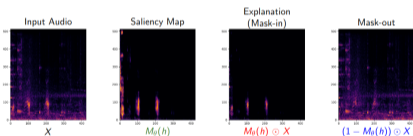
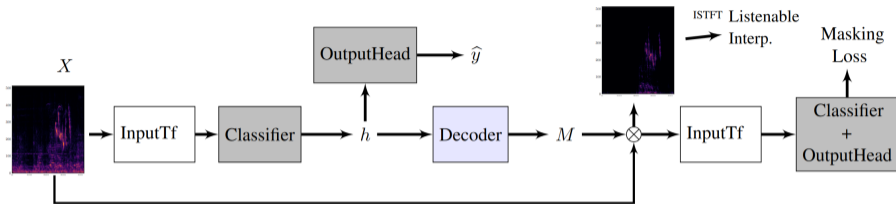
Input Audio



Explanation (Dog)



Listenable Maps for Audio Classifiers (L-MAC)



$$\min_{\theta} \underbrace{\lambda_{in} \mathcal{L}_{in}(\log f(M_{\theta}(h) \odot X), \hat{y})}_{\text{Mask-in}} - \underbrace{\lambda_{out} \mathcal{L}_{out}(\log f((1 - M_{\theta}(h)) \odot X), \hat{y})}_{\text{Mask-out}} + \underbrace{|M_{\theta}(h)|}_{\text{Mask Reg}}$$

	Metric	AI (↑)	AD (↓)	AG (↑)	FF (↑)	Fid-In (↑)	SPS (↑)	COMP (↓)
Listenable (STFT→Mel)	Saliency	0.00	15.79	0.00	0.05	0.07	0.39	5.48
	Smoothgrad	0.00	15.71	0.00	0.03	0.05	0.42	5.32
	IG	0.25	15.45	0.01	0.07	0.13	0.43	5.11
	GradCAM	8.50	10.11	1.47	0.17	0.33	0.34	5.64
	Guided GradCAM	0.00	15.61	0.00	0.05	0.06	0.44	5.12
	Guided Backprop	0.00	15.66	0.00	0.05	0.06	0.39	5.47
	L2I, RT=0.2	1.63	12.78	0.42	0.11	0.15	0.25	5.50
	SHAP	0.00	15.79	0.00	0.05	0.06	0.43	5.24
	L-MAC (ours)	36.25	1.15	23.50	0.20	0.42	0.47	4.71
	L-MAC, FT, $\lambda_p = 4$ (ours)	32.37	1.98	18.74	0.21	0.41	0.43	5.20
Not Listenable (Mel)	Saliency	0.00	15.81	0.00	0.10	0.07	0.39	4.53
	Smoothgrad	0.00	15.61	0.00	0.07	0.04	0.39	4.54
	IG	0.00	15.55	0.00	0.12	0.08	0.42	4.36
	GradCAM	7.00	10.93	1.04	0.17	0.29	0.34	4.72
	Guided GradCAM	0.125	15.40	6.67	0.08	0.07	0.45	4.17
	Guided Backprop	0.125	15.54	0.00	0.10	0.08	0.39	4.53
	SHAP	0.00	15.57	0.00	0.11	0.08	0.41	4.42
	L-MAC (ours)	35.63	1.59	24.28	0.22	0.42	0.45	4.11
	L-MAC (ours) FT, $\lambda_p = 4$	36.13	1.28	21.15	0.23	0.42	0.32	4.71

[F. Paissan, M. Ravanelli, C.Subakan; ICML 2024 (Oral)]

Table of Contents

Neural Network Explanations

Listenable Maps for Audio Classifiers

Faithfulness

Faithfulness in LALMs

Adversarial Interventions

Noise and Masking Interventions

Guided Masking

CoT Interventions

Grounding the LALM Reasoning

Measuring faithfulness and understandability

- **Faithfulness:** Measures importance of explanations for classifier decisions
 - ▶ L2I-Faithfulness

$$FF_n = p_{\hat{c}}(X_n) - p_{\hat{c}}(X_n - (X_n \odot M)),$$

Measuring faithfulness and understandability

- **Faithfulness:** Measures importance of explanations for classifier decisions
 - ▶ L2I-Faithfulness

$$FF_n = p_{\hat{c}}(X_n) - p_{\hat{c}}(X_n - (X_n \odot M)),$$



$P(x = \text{dog}) = 0.6$

$P(x = \text{dog}) = 0.01$

Measuring faithfulness and understandability

- **Faithfulness:** Measures importance of explanations for classifier decisions

- ▶ L2I-Faithfulness

$$FF_n = p_{\hat{c}}(X_n) - p_{\hat{c}}(X_n - (X_n \odot M)),$$

- ▶ Average-Increase

$$AI = \frac{1}{N} \sum_{n=1}^N [p_{\hat{c}}(X_n \odot M) > p_{\hat{c}}(X_n)] \cdot 100,$$



$P(x = \text{dog}) = 0.9$

$P(x = \text{dog}) = 0.6$

Measuring faithfulness and understandability

- **Faithfulness:** Measures importance of explanations for classifier decisions

- ▶ L2I-Faithfulness

$$FF_n = p_{\hat{c}}(X_n) - p_{\hat{c}}(X_n - (X_n \odot M)),$$

- ▶ Average-Increase

$$AI = \frac{1}{N} \sum_{n=1}^N [p_{\hat{c}}(X_n \odot M) > p_{\hat{c}}(X_n)] \cdot 100,$$

- ▶ Average-Gain

$$AG = \frac{1}{N} \sum_{n=1}^N \frac{\max(0, p_{\hat{c}}(X_n \odot M) - p_{\hat{c}}(X_n))}{1 - p_{\hat{c}}(X_n)} \cdot 100.$$

- ▶ Average-Drop

$$AD = \frac{1}{N} \sum_{n=1}^N \frac{\max(0, p_{\hat{c}}(X_n) - p_{\hat{c}}(X_n \odot M))}{p_{\hat{c}}(X_n)} \cdot 100.$$

- ▶ Input Fidelity

$$\text{Fid-In} = \frac{1}{N} \sum_{n=1}^N [\arg \max_c p_c(X_n) = \arg \max_{c'} p_{c'}(X_n \odot M)].$$

Table of Contents

Neural Network Explanations

- Listenable Maps for Audio Classifiers
- Faithfulness

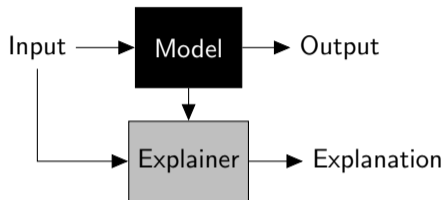
Faithfulness in LALMs

- Adversarial Interventions
- Noise and Masking Interventions
- Guided Masking
- CoT Interventions

Grounding the LALM Reasoning

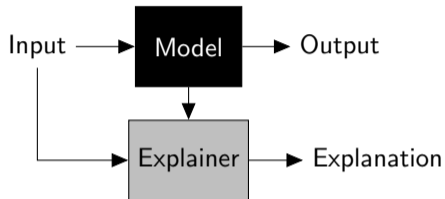
Faithfulness in LALMs

- The explanation pipeline we have been talking about



Faithfulness in LALMs

- The explanation pipeline we have been talking about



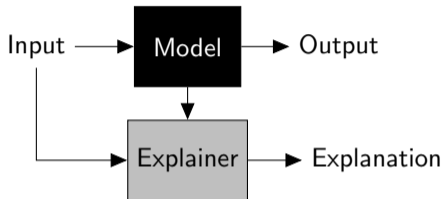
- More up-to-date explanation paradigm



- In the (multimodal) LLM realm, the explanations are the CoTs.

Faithfulness in LALMs

- The explanation pipeline we have been talking about

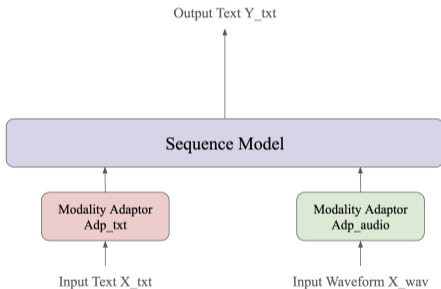


- More up-to-date explanation paradigm

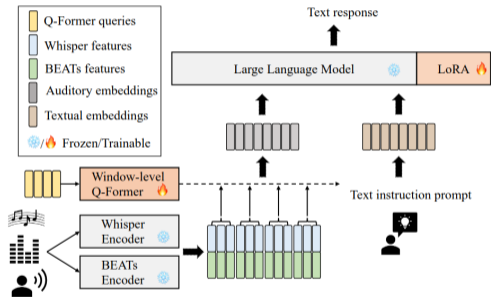
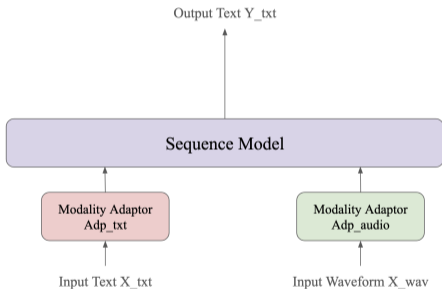


- In the (multimodal) LLM realm, the explanations are the CoTs.
- Or are they?

Large Audio Language Models



Large Audio Language Models



Can we use the CoTs as explanations?

Question: For the given audio, identify the background voices.

Choices: (A) Crowd (B) Solo Singer (C) Wind (D) Animal Sounds

Input Audio: Click to listen.

CoT: The audio contains multiple voices conversing simultaneously, which is characteristic of a crowd. There are no distinct sounds of a solo singer, wind, or animal sounds. The presence of multiple voices and the ambient noise suggest a social gathering or public space.

Answer: (A) Crowd

Can we use the CoTs as explanations?

Question: For the given audio, identify the background voices.

Choices: (A) Crowd (B) Solo Singer (C) Wind (D) Animal Sounds

Input Audio: Click to listen.

CoT: The audio contains multiple voices conversing simultaneously, which is characteristic of a crowd. There are no distinct sounds of a solo singer, wind, or animal sounds. The presence of multiple voices and the ambient noise suggest a social gathering or public space.

Answer: (A) Crowd

Is the CoT a faithful explanation for the decision?

Research Questions



[Mousavi et al, 2026, Investigating Faithfulness in Large Audio Language Models]

- Hallucination Free Listening

Research Questions



[Mousavi et al, 2026, Investigating Faithfulness in Large Audio Language Models]

- Hallucination Free Listening
- Wholistic Listening

Research Questions



[Mousavi et al, 2026, Investigating Faithfulness in Large Audio Language Models]

- Hallucination Free Listening
- Wholistic Listening
- Attentive Listening

Research Questions



[Mousavi et al, 2026, Investigating Faithfulness in Large Audio Language Models]

- Hallucination Free Listening
- Wholistic Listening
- Attentive Listening
- CoT Output Faithfulness

Research Questions



[Mousavi et al, 2026, Investigating Faithfulness in Large Audio Language Models]

- Hallucination Free Listening
 - Wholistic Listening
 - Attentive Listening
 - CoT Output Faithfulness
-
- We employ three different audio interventions and 4 different CoT interventions.

Audio Interventions

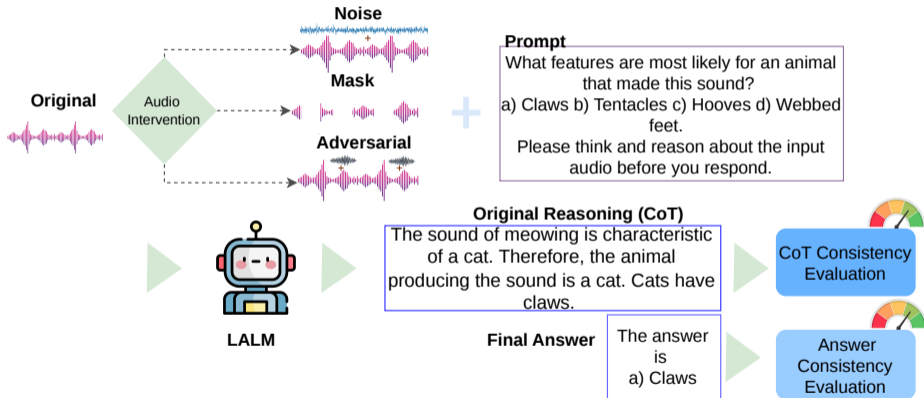


Table of Contents

Neural Network Explanations

- Listenable Maps for Audio Classifiers
- Faithfulness

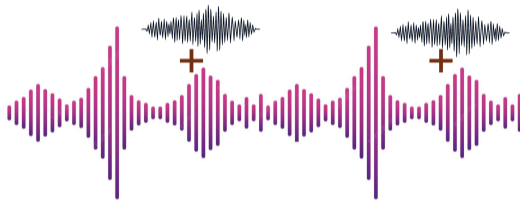
Faithfulness in LALMs

- Adversarial Interventions**
- Noise and Masking Interventions
- Guided Masking
- CoT Interventions

Grounding the LALM Reasoning

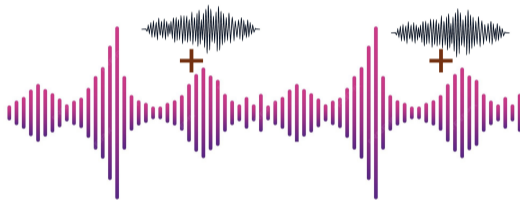
Adversarial Interventions

- We inject the wrong / correct answers.
- Pertinent research question: Attentive listening.
- We see if the model follows the prompt, or gets distracted.



Adversarial Interventions

- We inject the wrong / correct answers.
- Pertinent research question: Attentive listening.
- We see if the model follows the prompt, or gets distracted.



Original Audio

Correct Contamination

Wrong Contamination

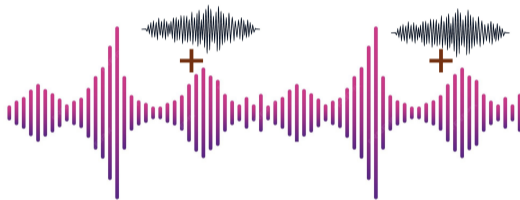
Original Audio

Correct Contamination

Wrong Contamination

Adversarial Interventions

- We inject the wrong / correct answers.
- Pertinent research question: Attentive listening.
- We see if the model follows the prompt, or gets distracted.



Original Audio

Correct Contamination

Wrong Contamination

Original Audio

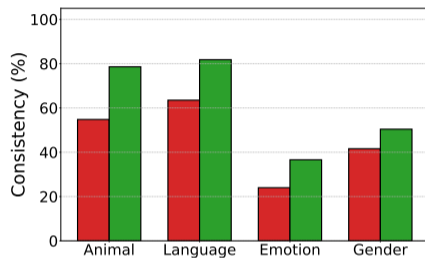
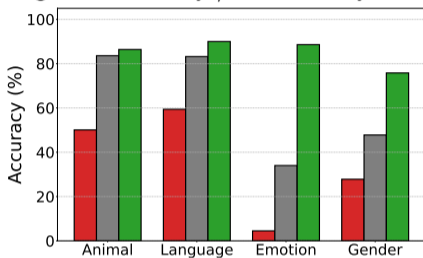
Correct Contamination

Wrong Contamination

More examples: [Website](#)

Adversarial Interventions - Consistency / Accuracy Change

■ The changes in Accuracy / Consistency, QWEN2.5 Omni:



■ Audio Flamingo 3:

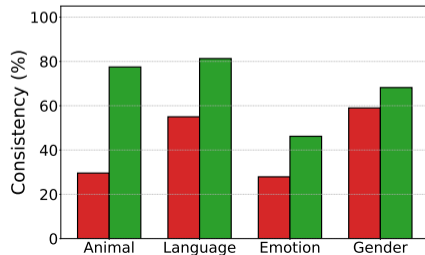
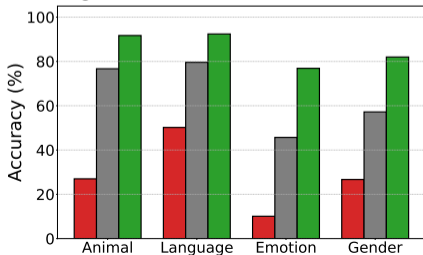


Table of Contents

Neural Network Explanations

- Listenable Maps for Audio Classifiers
- Faithfulness

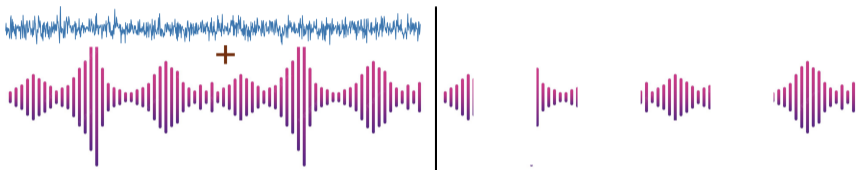
Faithfulness in LALMs

- Adversarial Interventions
- Noise and Masking Interventions
- Guided Masking
- CoT Interventions

Grounding the LALM Reasoning

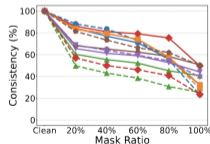
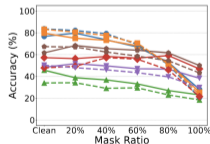
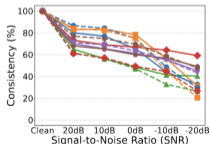
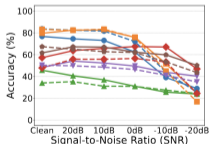
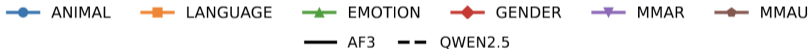
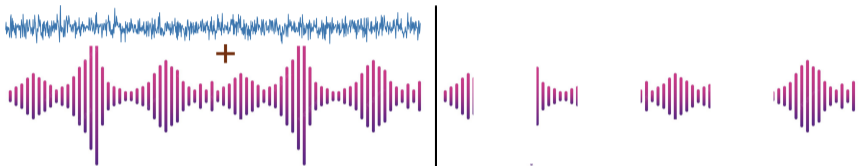
Noise and Masking Interventions

- We contaminate the signal with noise, and randomly drop chunks.
- **Research question:** Hallucination Free Listening, Wholistic Listening



Noise and Masking Interventions

- We contaminate the signal with noise, and randomly drop chunks.
- **Research question:** Hallucination Free Listening, Wholistic Listening



CoT consistency under noise and masking

Intervention	Model	Animal	Language	Gender	Emotion	MMAR	MMAU
Mask 100%	AF3	3.01	2.63	2.87	3.10	3.19	3.65
	Qwen	2.35	2.40	2.95	2.64	2.81	3.39
Mask 20%	AF3	4.60	4.32	4.35	4.04	3.97	4.41
	Qwen	4.68	4.62	3.89	3.45	4.01	4.45
-20dB SNR	AF3	2.89	2.54	3.17	3.13	3.09	3.57
	Qwen	2.45	2.27	2.70	2.61	2.82	3.22
20dB SNR	AF3	4.41	4.27	3.76	4.22	4.12	4.44
	Qwen	4.58	4.67	3.93	4.05	4.12	4.33

Hallucination Example

Question: Out of the provided options, which animal do you think is responsible for the sound captured in the provided audio recording?

Choices: (A) cow (B) rooster (C) sheep (D) dog

Original Audio: Click to listen.

Contaminated Audio: Click to listen.

Original CoT: The audio contains a distinct crowing sound. This sound is typically associated with a rooster. Frogs, dogs, and pigs do not make such a specific and recognizable crowing sound.

CoT after intervention: The audio contains a high-pitched, repetitive sound that is characteristic of a rooster's crowing. Frogs typically have croaking sounds, dogs bark, and pigs oink. The sound in the audio does not match the typical sounds of a frog, dog, or pig.

Judge explanation: Both reasonings identify a crowing sound in the audio and associate it with a rooster. They both eliminate frogs, dogs, and pigs based on their typical sounds, which do not match the crowing sound. The core logic and conclusion are identical, despite differences in wording and detail. **Similarity score:** = 5

Answer: (B) Rooster

Table of Contents

Neural Network Explanations

- Listenable Maps for Audio Classifiers
- Faithfulness

Faithfulness in LALMs

- Adversarial Interventions
- Noise and Masking Interventions
- Guided Masking**
- CoT Interventions

Grounding the LALM Reasoning

Guided Masking

[Mousavi et al. 2025, ICASSP 2025, What Are They Doing? Joint Audio-Speech Co-Reasoning]

Question: What are they possibly Doing?

Audio Description: Sound of seagulls and water lapping against a boat.

Speech Description: Speech of "Yeah! This one is much bigger!"

Guided Masking

[Mousavi et al. 2025, ICASSP 2025, What Are They Doing? Joint Audio-Speech Co-Reasoning]

Question: What are they possibly Doing?

Audio Description: Sound of seagulls and water lapping against a boat.

Speech Description: Speech of "Yeah! This one is much bigger!"

Audio-Oriented Answer: They are likely enjoying a day at the beach or sailing on a boat.

Speech-Oriented Answer: They are likely comparing the size of objects. Correct Answer: They are likely fishing on a boat.

Table of Contents

Neural Network Explanations

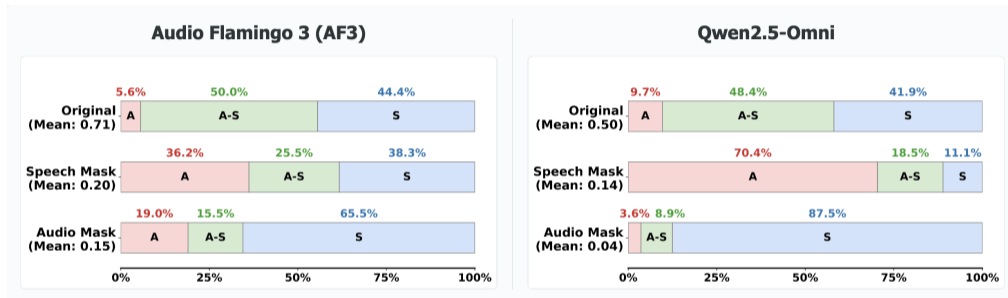
- Listenable Maps for Audio Classifiers
- Faithfulness

Faithfulness in LALMs

- Adversarial Interventions
- Noise and Masking Interventions
- Guided Masking
- CoT Interventions

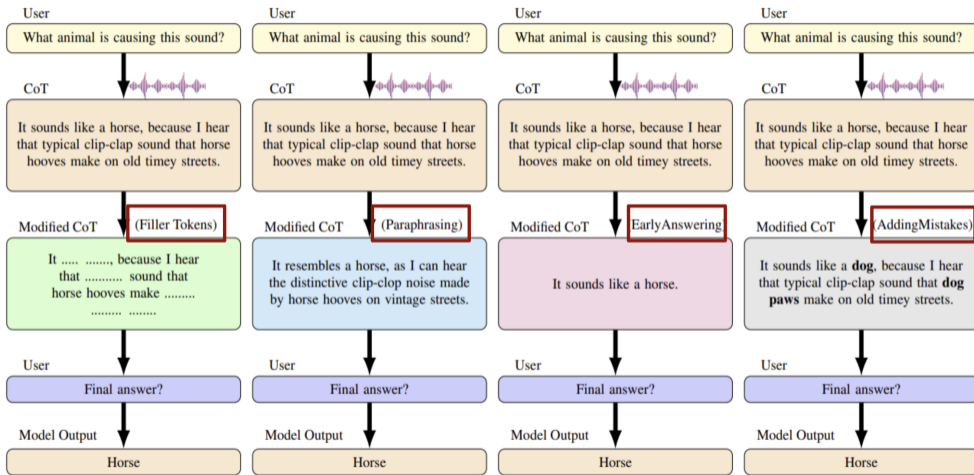
Grounding the LALM Reasoning

CoT Interventions

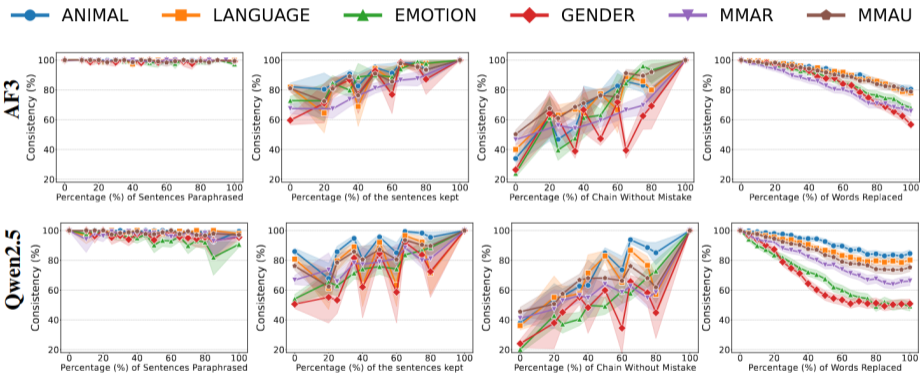


- **Wholistic listening:** Model plasticity is different for two models.
- **Hallucinations:** Models still attend to empty audio!

CoT interventions



CoT Interventions Results



Conclusions

- The Audio Ingestion does not seem to be very robust.
 - ▶ Hallucinations exist, the model does not always listen.
 - ▶ Models are not that robust, do not attentively listen, could be improved.
- CoT output faithfulness: Varies depending on the intervention.
 - ▶ Paraphrasing: Seems to be more robust than text models.
 - ▶ Early answering, Adding Mistakes: Exhibit more robust behavior.
 - ▶ Filler Tokens: Variations depending on the dataset, similar to text LLMs.

Table of Contents

Neural Network Explanations

- Listenable Maps for Audio Classifiers
- Faithfulness

Faithfulness in LALMs

- Adversarial Interventions
- Noise and Masking Interventions
- Guided Masking
- CoT Interventions

Grounding the LALM Reasoning

CoT with timestamps



A scenario of a customer dealing with an inefficient, frustrating bureaucratic system.

Q. Identify if any speaker experienced an emotion change in the given discussion.

Ours

From 5.98s to 14.80s, the first speaker expresses frustration, indicating that they have been seeking help for over an hour without success. This suggests that the first speaker is experiencing an emotion change, likely from neutral to frustrated.

The second speaker, from 15.94s to 19.42s, responds with a neutral tone, asking for clarification. There is no indication of an emotion change in the second speaker's response.

A. first speaker



Vanilla Reasoning

<think>

1. Speaker 1 starts with a polite tone, then becomes more frustrated and angry, and finally hopeful and relieved.
2. Speaker 2 starts with a neutral tone, then becomes inquisitive and slightly impatient, and finally reassuring and encouraging.

</think>

A. both speaker



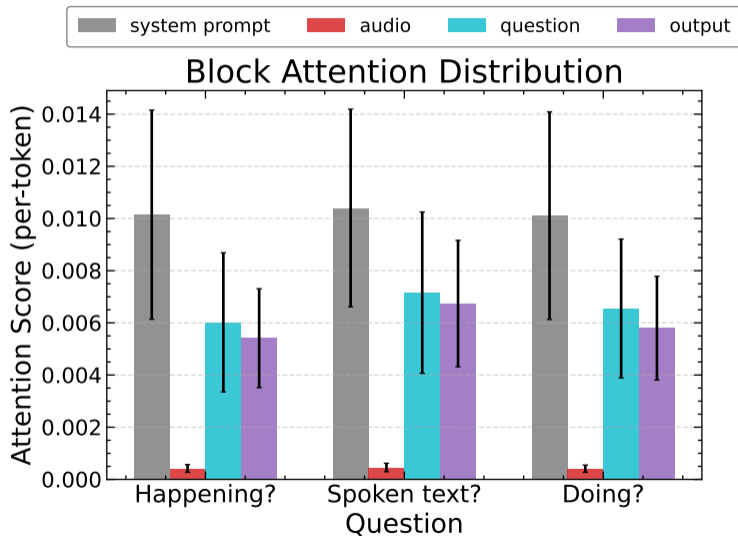
Zero-shot

A. none of the speakers



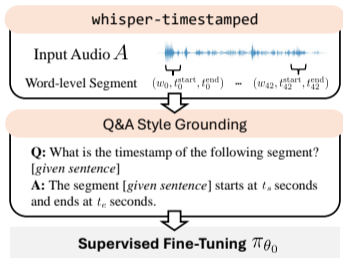
[Jeong et al., 2026, Listen First, Then Answer: Timestamp-Grouped Speech Reasoning]

Attention on Audio Tokens

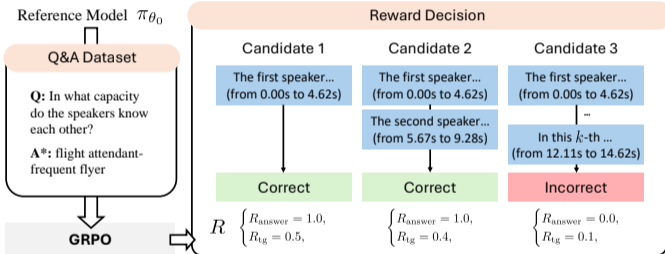


The method

Stage 1: Supervised Timestamp Alignment

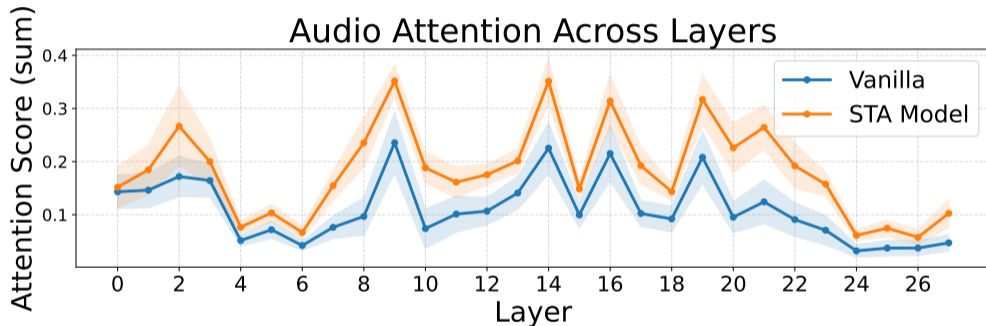


Stage 2: Timestamp-Grounded Reasoning



- We are supervising with sentence timestamps.

Attention after timestamps



Timestamp estimation accuracy

Table 1: *Effect of Sentence-level Timestamp Alignment.*

Model	STA	IoU \uparrow	F1 \uparrow	IoU ≥ 0.7 (%) \uparrow
Qwen-2.5-Omni [40]	\times	0.2324	0.3074	11.32
	\checkmark	0.7189	0.7675	69.79
AudioFlamingo3 [17]	\times	0.2838	0.3592	11.45
	\checkmark	0.5439	0.6012	47.95
Gemini-2.5-Flash [44]	–	0.7356	0.8141	71.08

How well does it perform?

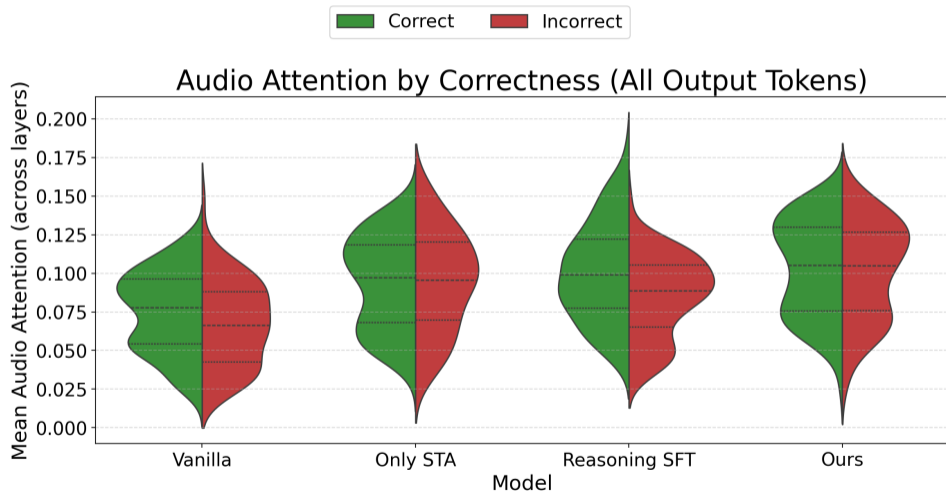
Methods	Size	MMAU-mini Speech (%)	MMAR-Speech (%)	AIR-Bench SER (%)	AIR-Bench SNV (%)	AIR-Bench SIC (%)	MELD (%)
<i>Proprietary Models</i>							
Gemini-2.5-Flash	–	75.08	72.41	56.4	68.5	83.6	61.5
GPT-4o Audio	–	66.67	20.41	51.2	61.6	89.3	62.5
<i>Open-source Models</i>							
SALMONN	7B	26.43	24.35	29.0	34.3	42.3	37.2
Audio Flamingo 3	7B	66.37	57.48	59.5	76.8	79.6	58.5
<i>Audio Reasoning Methods</i>							
Audio-CoT	8.4B	55.26	54.01	–	–	–	–
Audio-Reasoner	8.4B	66.07	32.99	60.5	56.3	88.1	63.2
Audio-Thinker	8.4B	73.37	64.29	56.2	67.5	–	–
<i>Our Ablation Variants</i>							
Qwen2.5-Omni (baseline)	7B	70.60	59.86	60.2	63.9	83.5	60.3
+ Only STA	7B	71.37	61.22	59.5	66.0	84.3	62.8
+ Reasoning SFT	7B	74.47	62.93	58.5	68.1	85.0	61.8
Ours (Full)	7B	74.47	64.63	62.5	70.4	89.3	64.6

Let's listen to some examples

- Our website.



Final Attention on Audio



Conclusions

- The proposed time stamp ground improves the attention on audio tokens.
- The model outputs accurate time-stamps, provides a listenable reasoning.
- The performance is improved over baseline models.

Students leading

Pooneh Mousavi



Lovenya Jain



Jihoon Jeong (Huni) Francesco Paissan



Thanks for listening!

