Audio Entailment: Assessing Deductive Reasoning for Audio Understanding

Soham Deshmukh[1,2], **Shuo Han**[1], **Hazim Bukhari**[1], **Benjamin Elizalde**[2], **Hannes Gamper**[3], **Rita Singh**[1], **Bhiksha Raj**[1]

[1]Carnegie Mellon University
[2]Microsoft
[3]Microsoft Research
{sdeshmuk, shuohan, hbukhari}@andrew.cmu.edu

Abstract

Recent literature uses language to build foundation models for audio. These Audio–Language Models (ALMs) are trained on a vast number of audio–text pairs and show remarkable performance in tasks including Text-to-Audio Retrieval, Captioning, and Question Answering. However, their ability to engage in more complex open-ended tasks, like Interactive Question-Answering, requires proficiency in logical reasoning—a skill not yet benchmarked. We introduce the novel task of Audio Entailment to evaluate an ALM's deductive reasoning ability. This task assesses whether a text description (hypothesis) of audio content can be deduced from an audio recording (premise), with potential conclusions being entailment, neutral, or contradiction, depending on the sufficiency of the evidence. We create two datasets for this task with audio recordings sourced from two audio captioning datasets—AudioCaps and Clotho—and hypotheses generated using Large Language Models (LLMs). We benchmark state-of-the-art ALMs and find deficiencies in logical reasoning with both zero-shot and linear probe evaluations. Finally, we propose "caption-before-reason", an intermediate step of captioning that improves the Zero-Shot and linear-probe performance of ALMs by an absolute 6% and 3%, respectively.

Datasets — https://github.com/microsoft/AudioEntailment

1 Introduction

Recent literature uses language to build foundation models for audio. These models, referred to as Audio–Language Models, are trained on millions of audio–text pairs using either Contrastive Learning (e.g., CLAP (Elizalde et al. 2023; Wu et al. 2023)) or Next-Token Prediction (e.g., Pengi (Deshmukh et al. 2023), Qwen-Audio (Chu et al. 2023)). Once trained, ALMs can perform multiple tasks grounded in audio and user-provided instructions, for example text-to-audio retrieval, captioning, question-answering, and text-to-audio generation. Owing to their performance, support for various tasks, and inherent ease-of-use, ALMs are being extensively used across various scenarios.

ALMs have achieved state-of-the-art (SoTA) performance on close-ended tasks like Classification and Retrieval, beating Self-Supervised Learning (SSL) models as well as Supervised models. The latest ALMs efforts (Chu et al. 2023; Gong et al. 2023a; Tang et al. 2024) focus on improving open-ended text generation. The task (Deshmukh et al. 2023) consists of generating free-form text, given an audio and a text input, and has flexibility in the correctness of the output. For instance, an audio recording labeled as "dog barking" can be identified by the ALMs as "canine barking" and still be marked as correct. The open-ended text generation for ALMs usually takes the form of interactive question-answering with the user. From a Machine Learning perspective, one can think of a model performing different tasks of Audio Captioning, Audio Question Answering, Audio Dialogues, and Reasoning, to enable interactive Question-Answering. To generate natural and accurate responses, the ALMs should have learned to think step-by-step, utilize the learned real-world knowledge, and have the ability to ask follow-up questions for clarifications about the acoustic content. ALMs are evaluated on such abilities through Audio Question Answering tasks. Although the performance has been promising, ALMs do not perform well on interactive Question-Answering. Hence, we introduce a new direction to evaluate a specific type of reasoning of ALMs called Logical Reasoning.

Logical Reasoning (Copi, Cohen, and McMahon 2016) can be defined in the context of a premise and a hypothesis. To perform Logical Reasoning, one needs a comprehension of premises, the relationships among premises, and then use rigorous methods to infer conclusions that are implied by the premises and relations. Deductive reasoning, a form of Logical Reasoning, is useful where the premises are known to be true, as it allows for drawing specific conclusions from general principles. Deductive reasoning in audio perception involves a "top-down" approach, where one begins with hearing an audio and then determines if a logical conclusion can be drawn. For instance, an audio contains a dog barking and children playing. The hypothesis is "children playing in the park with a dog barking nearby." Thus, we can conclude the hypothesis is plausible, as parks can be associated with these sounds. Evaluating deductive reasoning also helps in identifying audio hallucinations. They may manifest in two ways: (1) Inferred Cues: The model generates cues not present in the audio input, such as introducing audio events that were neither mentioned nor implied. (2) Contextual Events: The model relies on contextual assumptions rather than audio evidence, for example, interpreting a sound as "dog barking"

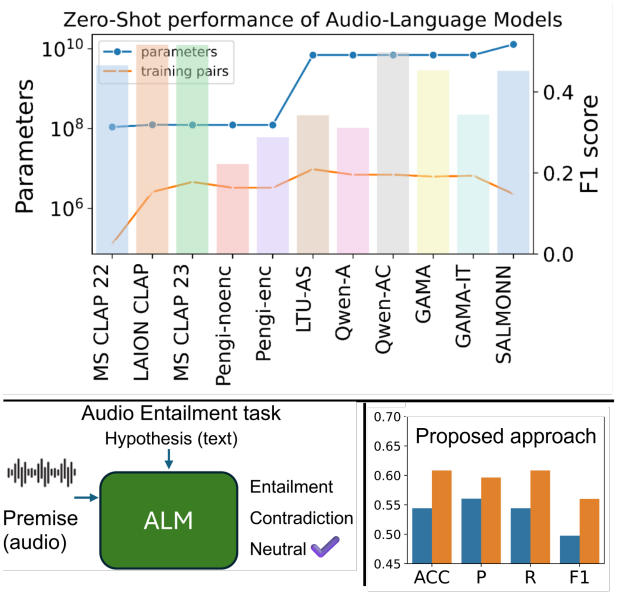

Figure 1: (Bottom left) Audio-Language Models have to infer Entailment, Neutral, or Contradiction from an audio premise $\mathcal{P}$ and a textual hypothesis $\mathcal{H}_*$. (Top) The highest performing Zero-Shot inference (or classification) is 57% F1 from LAION CLAP. (Bottom right) Our proposed method, combining MS CLAP 23 and a captioning step, enhances performance by an absolute 3% F1.

because the word "dog" is usually followed by "barking", while the audio more accurately suggests "whimpering" or other actions. By benchmarking ALMs for deductive reasoning, we can uncover audio hallucinations.

In this work, we study Logical Reasoning for ALMs. Our contributions are:

- We introduce the task of *Audio Entailment* to test the Deductive Reasoning ability of ALMs. The task determines if a textual hypothesis $\mathcal{H}$ can be concluded from an audio premise $\mathcal{P}$. The conclusion can be entailment, neutral, or contradiction based on the evidence. We created two datasets, ACE and CLE, where Hypotheses were first generated by GPT-4 and then verified and corrected by human annotators. This two-step process enhances the quality of the datasets, which will be publicly released.
- We benchmark SoTA ALMs, showing they have limited deductive reasoning. We test both contrastive and next-token prediction ALMs in Zero-Shot and linear-probe setups and highlight ways to enhance audio-grounded reasoning.
- Based on our findings, we propose "caption-before-reason" which performs intermediate captioning before reasoning, improving zero-shot and linear-probe performance by an absolute 6% and 3%, respectively.

2 Related Work

Audio-Language Models. The early models focused on close-ended tasks. For example, CLAP (Elizalde, Deshmukh, and Wang 2024; Wu et al. 2023; Dhamyal et al. 2024) is contrastively trained on millions of audio-text pairs and learns multimodal representations that can be used for close-ended tasks like Zero-Shot classification and retrieval. With the success of CLAP, later ALMs focused on tackling open-ended tasks, like Audio Captioning or Audio Question Answering (AQA). For example, Pengi (Deshmukh et al. 2023) and LTU (Gong et al. 2023b) concurrently framed all audio tasks as audio-and-text input to text output tasks. In terms of architecture, Pengi and LTU jointly train an audio encoder with a frozen or near-frozen LLM. Each is capable of producing text based on audio inputs and text prompts. The subsequent generation of ALMs focus on performing joint speech–audio understanding and utilize larger training data and LLMs. For example, Qwen-Audio (Chu et al. 2023), LTU-AS (Gong et al. 2023a), GAMA (Ghosh et al. 2024b), AudioFlamingo (Kong et al. 2024) and SALMONN (Tang et al. 2024) beat existing ALMs on 30 different tasks, each showcasing unique strengths and weaknesses.

Audio Question Answering (AQA). The task involves analyzing an audio signal and a question to provide accurate answers. There are two AQA datasets in the literature to train and test ALMs. (1) ClothoAQA (Lipping et al. 2022) is a crowdsourced dataset consisting of 1991 audio files, selected from the Clotho dataset (Drossos, Lipping, and Virtanen 2020). It includes a set of six different questions and corresponding answers for each audio file, which were collected through crowdsourcing using Amazon Mechanical Turk. (2) OpenAQA (Gong et al. 2023b) combines 5 different datasets from the literature and converts them into a triplet format of: audio input, text prompt, and text output. It includes 1.9M close-ended questions and 3.7M open-ended questions generated with the help of GPT-3.5-Turbo (Brown et al. 2020). However, neither dataset evaluates deductive Reasoning.

Text and Visual Entailment. Natural Language Inference (MacCartney 2009; Dagan, Glickman, and Magnini 2005), also known as Textual Entailment, is a concept in Natural Language Processing that involves determining the relationship between two text fragments. The relationship is directional and holds whenever the truth of one text fragment (the premise) follows from another text (the hypothesis). For example, if the premise is "The cat sat on the mat", and the hypothesis is "There is a cat on a mat", then we can infer that the hypothesis is true given the premise. Visual Entailment (Xie et al. 2019; Do et al. 2020) extends this to the vision domain where the image is the premise and a text fragment is the hypothesis. The task is to predict whether the image semantically entails the text. This type of reasoning is shown to be crucial for fine-grained image understanding (Thomas, Zhang, and Chang 2022). Recent research has identified perception gaps in reasoning (Ghosh et al. 2024a).

3 Audio Entailment

Entailment (Routley and Meyer 1973; Anderson, Belnap Jr, and Dunn 2017) holds when there is a directional relationship between the premise ($\mathcal{P}$) and hypothesis ($\mathcal{H}$). Specifically, for our work, we use a relaxed definition: "p entails h" ($\mathcal{P} \Rightarrow \mathcal{H}$) if, typically, ***a human observing*** $\mathcal{P}$ would infer that $\mathcal{H}$ is most likely true. This relation is directional,

meaning that even if $\mathcal{P} \Rightarrow \mathcal{H}$, the reverse $\mathcal{H} \Rightarrow \mathcal{P}$ is uncertain. Entailment helps determine whether a hypothesis logically follows from the premise, allowing us to infer relationships between premise and hypothesis fragments. We consider various definitions of audio entailment, and specifically choose a definition based on inferential analysis (details in Appendix).

In Audio Entailment, the premise $\mathcal{P}$ is audio recorded in-the-wild and the hypothesis $\mathcal{H}$ is a natural language description. The aim of the Audio Entailment task is to determine if the hypothesis $\mathcal{H}$ can be concluded by a human listening to the audio recording premise P. This leads us to the following three scenarios (Fig. 2):

- Entailment is determined when the audio recording $\mathcal{P}$ contains sufficient evidence to affirm the truth of the hypothesis $\mathcal{H}$.
- Neutral holds when the audio recording $\mathcal{P}$ does not provide enough information to either confirm or deny the hypothesis $\mathcal{H}$. Simply put, while $\mathcal{H}$ may be true, it cannot be substantiated solely from the audio recording $\mathcal{P}$.
- Contradiction is determined when the audio recording $\mathcal{P}$ offers substantial evidence to deduce that the hypothesis $\mathcal{H}$ is false.

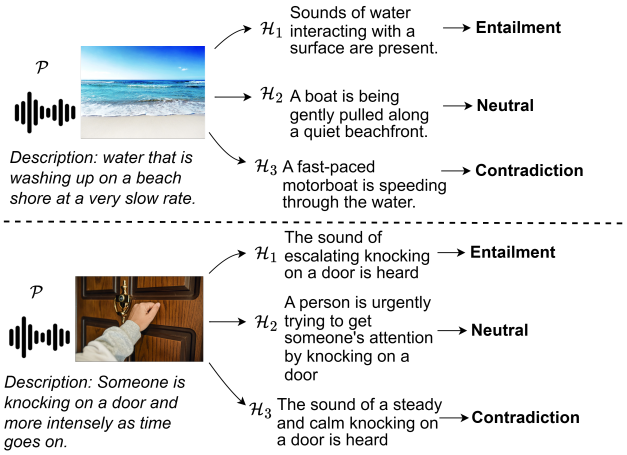

Figure 2: The figure shows two examples of the Audio Entailment task. The image and *Description* are provided here for illustration purposes and are not part of the task. The premise $\mathcal{P}$ consists of an audio recording and a hypothesis $\mathcal{H}_*$. Given the premise, Audio Entailment is determined for H_1, Neutral for H_2, and Contradiction for H_3 respectively.

3.1 Audio Entailment as a Classification Task

We formulate the Audio Entailment task as a classification task. The input consists of $\{a_i, h_i\}$, with audio premise a_i and hypothesis h_i, and the target is to predict $\{c\}$, where $c \in \{\text{entailment, neutral, contradiction}\}$. To make an accurate prediction c, the model has to understand the relation between a_i and h_i, enforcing and verifying a step of logical reasoning.

4 Audio Entailment Datasets

In this section, we describe the creation of AudioCaps Entailment (ACE) and Clotho Entailment (CLE).

4.1 Audio Premise

The premise $\mathcal{P}$ for Audio Entailment is a real-world audio recording. We source audio files and their corresponding natural language annotations from two Audio Captioning datasets, AudioCaps (Kim et al. 2019) and Clotho (Drossos, Lipping, and Virtanen 2020).

AudioCaps. The AudioCaps dataset comprises 46,000 audio samples sourced from AudioSet, each labeled with a single caption. These captions were collected through the Amazon Mechanical Turk (AMT) crowdsourcing platform, complemented by automated checks for the quality of annotations. Annotators were given the word labels from AudioSet and had access to the corresponding videos for the audio clips they were annotating. It should be noted that providing annotators access to visuals may introduce bias if annotators focus on the visual elements rather than the auditory ones. Furthermore, limiting the data to a single caption for each file hinders the ability of ALMs to learn and assess a wide range of descriptions. Finally, as AudioCaps derives its content from YouTube, there has been a gradual loss of videos over time, resulting in the unavailability of certain audio files. To counteract some of these limitations, we rely on Clotho as an additional dataset.

Clotho. The Clotho audio collection is obtained from the Freesound platform. This platform enables individuals to share their audio recordings and accompany them with descriptions. These recordings range in length from 15 to 30 seconds. For each audio clip, there are five captions, each containing 8 to 20 words. These captions are gathered using AMT, following a detailed protocol for crowdsourcing audio captions to promote variety and minimize grammatical mistakes. The annotators had access solely to the audio tracks, without any additional context such as video or textual tags, during the annotation process.

Other existing datasets (SoundDescs (Koepke et al. 2022), MACs (Martín-Morató and Mesaros 2021) and WT5K (Deshmukh, Elizalde, and Wang 2023)) do not contain human annotations and are therefore not considered for building the first version of the audio entailment dataset.

4.2 Hypothesis

From Clotho and AudioCaps, we obtain audio recordings and the natural language description of the audio. The natural language descriptions in these sets are created by humans, and aim to be as descriptive as possible, often including the source of the sound, the action taking place, and any additional context that can be inferred from the audio. For example, a caption will not only state "dog barking" but expand to "a dog barking loudly in the distance, with the sound of traffic in the background," giving a more complete picture of the auditory scene. Hence, the language description can serve as a succinct substitute for a typical human description of the audio recording. This text-based version allows for the generation of hypotheses through the use of an LLM.

Sample 1. A person is flipping quickly the pages of a book.	
[Entailment]	A person is moving the pages of a book or paper.
[Neutral]	A person is organizing documents and occasionally flipping through pages.
[Contradiction]	A person is typing on a computer keyboard.
Sample 2. A variety of birds chirping and singing and shoes with a hard sole moving along a hard path.	
[Entailment]	Birds are chirping outdoors while someone with hard-soled shoes walks on a hard surface.
[Neutral]	A child is playing outside where birds are singing and someone is walking on a cobblestone path nearby.
[Contradiction]	A choir is performing in a concert hall.
Sample 3. Many people are speaking simultaneously in a public place before a man hollers out something.	
[Entailment]	A noisy indoor environment with multiple conversations happening and an occasional shout from an individual.
[Neutral]	Customers are chatting in a crowded cafe as a barista announces a ready order.
[Contradiction]	A quiet library setting with people whispering and no sudden loud voices.

Table 1: Audio Entailment examples from the AudioCaps Entailment and Clotho Entailment datasets we introduce in this study.

Our approach consists of two steps, hypothesis generation and hypothesis verification.

Hypothesis Generation. LLMs are known to exhibit reasoning abilities when they are *sufficiently large* (Huang and Chang 2023) (Wei et al. 2022b). For instance, using techniques including a "chain of thought" approach, such as reasoning examples, or even a straightforward prompt like "Let's consider this one step at a time," these models can tackle queries by outlining clear, logical steps. This method has been demonstrated in studies (Wei et al. 2022a; Kojima et al. 2022) and enables logical deduction like "if all birds have wings and all wings enable flight, then it logically follows that all birds can fly". Hence, we use a closed-source (GPT4) and an opensource LLM (Llama3) to generate potential hypotheses for the three cases, entailment, neutral, and contradiction. We experimented with various prompting techniques, and identified three primary strategies that yielded results anchored in audio descriptions: (1) Directing the LLM to explicitly utilize knowledge from audio, acoustics, and psychoacoustics for hypothesis generation. (2) Incorporating hard examples within the prompts to obtain a better hypothesis for the neutral case. (3) Explicit instructions to avoid negations and "easy" neutral and contradiction examples. The exact prompt used is described in Appendix.

Hypothesis Verification. Our rationale for employing LLMs to create hypotheses is based on the assumption that "language descriptions can act as a compact and precise alternative to a typical human description of the audio recordings," although this may not be reliable if errors occur in the annotator's audio descriptions. To counteract this, we employ five distinct descriptions from separate annotators for each audio file to formulate three hypotheses. Providing the LLM with five varied descriptions ensures that it capitalizes on the commonalities among them, thereby minimizing the impact of human annotation errors on hypothesis generation. Subsequently, once the LLM generates hypotheses for each scenario—entailment, neutrality, and contradiction—— we engage human annotators to either reject or validate these hypotheses. Should a hypothesis be rejected, the annotators will listen to the audio and propose an alternative hypothesis. This verification step ensures the Audio Entailment dataset is devoid of problematic hypotheses. Our two-step method——leveraging LLM for initial hypothesis generation followed by human verification and correction of challenging hypotheses——provides a balance between cost and time efficiency.

Data	Split	Dur. [hrs]	$\mathcal{H}$	Median [chars]	Max [chars]	Vocab. [words]
CLE	train	23.98	3839	68	195	4678
CLE	val	6.56	1045	69	208	2828
CLE	test	6.50	1045	67	192	2759
ACE	test	2.63	4785	57	207	3901

Table 2: Statistics of AudioCaps Entailment (ACE) and Clotho Entailment (CLE) (cf. Sec 4.3).

4.3 AudioCaps and Clotho Entailment

The Audio Entailment dataset consist of $\{a_i, h_i, c_i\}$ triplets, with audio premise a_i, hypothesis h_i, and the target c_i where $c \in \{\text{entailment, neutral, contradiction}\}$. We create this dataset for AudioCaps (Kim et al. 2019) and Clotho (Drossos, Lipping, and Virtanen 2020) using steps described in Sec. 4.1 and Sec. 4.2. The dataset statistics and samples from the dataset are shown in Table 2 and Table 1 respectively. We generate hypotheses for all sets of Clotho and restrict to only the test set of AudioCaps. The train set of AudioCaps has only one caption per recording and leads to generated hypotheses not aligned with the audio content. Hence, we only generate hypotheses for AudioCaps test set which has five captions per audio recording. To calculate median and max number of characters per hypothesis in Table 2 we preprocess the hypotheses $\mathcal{H}$ by dividing them into words, converting all letters to lowercase, and removing punctuation. The total vocabulary size per set is in the last column. Duration of the total audio is in hours. An analysis of the audio content in the proposed datasets can be found in the Appendix. We conduct experiments using ACE and CLE in Section 5 on 80GB A100 GPU.

Data	ALM	AE (params)	LLM (params)	ACC↑	P↑	R↑	F1↑	EACC↑	NACC↑	CACC↑
CLE	MS CLAP 22	CNN14 (80M)	BERT (110M)	0.4590	0.5499	0.459	0.4656	0.6000	0.4029	0.3742
CLE	LAION CLAP	HTSAT (31M)	RoBERTa (125M)	0.5113	**0.5544**	0.5113	**0.5161**	**0.6679**	0.3646	0.5014
CLE	MS CLAP 23	HTSAT (31M)	GPT2 (124M)	**0.5164**	0.5155	**0.5163**	0.5159	0.4153	**0.4038**	**0.7301**
ACE	MS CLAP 22	CNN14 (80M)	BERT (110M)	0.4334	0.4435	0.4334	0.4332	0.4332	0.5641	0.4508
ACE	LAION CLAP	HTSAT (31M)	RoBERTa (125M)	**0.5872**	**0.5767**	**0.5872**	**0.5693**	0.2867	**0.5900**	**0.8848**
ACE	MS CLAP 23	HTSAT (31M)	GPT2 (124M)	0.4860	0.4678	0.4860	0.4656	**0.4880**	0.2002	0.7699

Table 3: Zero-Shot performance of Contrastive Audio Language Models on Audio Entailment.

5 Deductive Reasoning With ALMs

This section benchmarks the deductive reasoning capabilities of SoTA ALMs. The deductive reasoning task is framed as a 3-way classification task, and hence we use classification metrics including accuracy, precision, recall, and F1.

5.1 Audio-Language Models

Recent ALMs in the literature can be broadly divided into (a) contrastive and (b) next-token prediction.

Contrastive ALMs use a two-tower structure consisting of audio and text encoders. The two branches are trained using contrastive learning and learn a joint audio–text multimodal space. After training, the model can be used for Zero-Shot inference for close-ended classification and retrieval tasks. Examples are MS CLAP (Elizalde et al. 2023) and LAION CLAP (Wu et al. 2023). In the case of contrastive ALMs, the audio premise and text hypothesis are encoded by the audio and text branch, respectively. We compute the dot product between the audio and text embeddings to obtain a score. We use non-overlapping similarity thresholds to predict the three classes entailment, neutral, and contradiction. The specifics of the thresholding method can be found in the Appendix. Classifying predictions into three categories via score thresholds eliminates the need for post-processing.

Next-token prediction ALMs take an audio recording and text as input and generate free-form text as output. The input audio is converted into a sequence of continuous embeddings using an audio encoder and is used to prompt a frozen or near-frozen (LoRA) LLM. Examples are Pengi (Deshmukh et al. 2023), LTU-AS (Gong et al. 2023a), Qwen-Audio (Chu et al. 2023). In this case, the audio premise becomes the audio input and the text hypothesis becomes the text prompt. The output of next-token ALMs are complex descriptions. Therefore, we use an LLM to classify the ALM descriptions into 3 classes. The text prompt used for each ALM and details on LLM-based evaluation is available in Appendix. Results are 5-run averages.

5.2 Zero-Shot Performance on Audio Entailment

The Zero-Shot performance of contrastive models is summarized in Table 3 and Next-token results are reported in Table 4. We make the following observations: (1) **Larger language models improve deductive reasoning but are challenging to ground in audio.** Among the next-token prediction ALMs, Pengi uses GPT2-base, a 128M parameter decoder while the rest use 7B LLM or larger as the decoder. We observe that the larger the LLM and its pretraining, the better the F1 score on the audio entailment task. For example, GAMA outperforms LTU-AS. Both models use largely the same training data based on OpenAQA, but GAMA uses Llama2 7B instead of Vicuna (based on Llama 7B) used by LTU-AS. However, with larger language models and their pretraining, we observe models hallucinating responses more; minor changes in prompt lead to ALMs hallucinating audio events and completely changing their deduction. For example, changing stopwords like "it" to "the" in the prompts of SALMONN and GAMA leads to them changing the deduction from contradiction to "yes, the audio events are present in the clip and hence it is true". Without any instruction-based fine-tuning, the models rely heavily on language statistics without aligning with audio or human intent. For example, Qwen Audio uses Qwen-7B as the initialization of the LLM, and Whisper-large-v2 as the initialization of the audio encoder. The Qwen-Audio Chat version utilizes the base Qwen-Audio and undergoes instruction-based fine-tuning to improve the ability of the model to align with human intent. We observe minor hallucinations with Qwen-Audio Chat version compared to other ALMs. (2) **Training ALMs to predict uncertainty improves their ability to detect plausible scenarios.** All evaluated next-token prediction ALMs have the lowest accuracy for determining whether the hypothesis is plausible given the audio premise, compared to entailment or contradiction. We observe models like Pengi, Qwen-Audio are more likely to predict entailment instead of any other response. However, GAMA and LTU-AS are the two-top performing models in determining if the hypothesis is plausible given the audio premise. This can be attributed to the training recipe used for the model. GAMA and LTU-AS are trained on more than 3.7M QA pairs generated using GPT-3.5 Turbo, and about 6.5% contain "I don't know" or "cannot answer due to insufficient information". By training on these pairs, the authors aim to reduce model hallucinations and avoid answering questions that cannot be addressed solely by audio. For the task of deductive reasoning, the model can now use this ability to better predict if the audio recording does not provide sufficient evidence to either confirm or deny the hypothesis. However, this increase in detecting neutral is only achieved when the prompt matches the training data (details in Appendix). Also, the increase in detecting neutral comes at the cost of entailment accuracy, where the model is more likely to say "I cannot say" even if the audio has sufficient evidence to determine the hypothesis is true. Our proposed "caption-

Data	ALM	AE (params)	LLM (param)	ACC↑	P↑	R↑	F1↑	EACC↑	NACC↑	CACC↑
CLE	Pengi-noenc	HTSAT (31M)	GPT2 (124M)	0.2781	0.1843	0.2781	0.2216	0.4967	0.0000	0.3378
CLE	Pengi-enc	HTSAT (31M)	GPT2 (124M)	0.3726	0.2465	0.3726	0.2888	0.7541	0.0000	0.3636
CLE	LTU-AS	Whisper-L (640M)	Vicuna (7B)	0.3681	0.3737	0.3681	0.3420	0.6278	0.3187	0.1579
CLE	Qwen-A	Whisper-L (640M)	Qwen (7B)	0.3620	0.4012	0.3620	0.3117	0.7675	0.1388	0.1799
CLE	Qwen-AC	Whisper-L (640M)	Qwen (7B)	**0.5442**	0.5604	**0.5442**	**0.4975**	**0.9024**	0.1569	0.5732
CLE	GAMA	CAV-MAE (85M)	LLaMA2 (7B)	0.4826	**0.6151**	0.4826	0.4534	0.8144	**0.4124**	0.2211
CLE	GAMA-IT	CAV-MAE (85M)	LLaMA2 (7B)	0.3974	0.5604	0.3974	0.3433	0.7923	0.2947	0.1053
CLE	SALMONN	Combined* (730M)	Vicuna (13B)	0.5222	0.5054	0.5222	0.4515	0.6775	0.0708	**0.8182**
ACE	Pengi-noenc	HTSAT (31M)	GPT2 (124M)	0.2629	0.1699	0.2629	0.2045	0.5312	0.0000	0.2575
ACE	Pengi-enc	HTSAT (31M)	GPT2 (124M)	0.3867	0.2558	0.3867	0.3039	0.7335	0.0000	0.4265
ACE	LTU-AS	Whisper-L (640M)	Vicuna (7B)	0.3633	0.3772	0.3633	0.3334	0.6702	0.2435	0.1762
ACE	Qwen-A	Whisper-L (640M)	Qwen (7B)	0.3563	0.3562	0.3563	0.3219	0.6669	0.1323	0.2696
ACE	Qwen-AC	Whisper-L (640M)	Qwen (7B)	0.5216	0.5669	0.5216	0.4918	**0.9300**	0.2821	0.3528
ACE	GAMA	CAV-MAE (85M)	LLaMA2 (7B)	0.5248	**0.6531**	0.5248	**0.4933**	0.7827	**0.5885**	0.2031
ACE	GAMA-IT	CAV-MAE (85M)	LLaMA2 (7B)	0.4167	0.5672	0.4167	0.3828	0.7852	0.2696	0.1954
ACE	SALMONN	Combined* (730M)	Vicuna (13B)	**0.5622**	0.5551	**0.5622**	0.4826	0.7114	0.0698	**0.9055**

Table 4: Zero-Shot performance of Next-token prediction Audio Language Models on Audio Entailment. The combined* Audio Encoder (AE) indicates a concatenation of Whisper-Large and BEATs audio encoder.

before-reason" method improves this behaviour (Sec. 5.4) (3) **Contrastive models are competitive on the task of deductive reasoning.** The contrastive models perform comparably to the next-token prediction models on the task of deductive reasoning. One main reason is that contrastive models include both audio and text encoders that capture sentence-level information, making them ideal for classification tasks. Second, Contrastive models need a classification threshold, unlike next-token prediction models that give direct answers. Tuning this threshold can improve their performance. We use non-overlapping thresholds (details in Appendix) to test the natural separability of the latent space of these models. Even with non-overlapping linearly increasing thresholds, we see F1 scores of around 50%. This indicates that the CLAP similarity score, which is the distance between the audio and text embeddings in the latent space, changes linearly with the alignment of the hypothesis with the audio premise. This makes contrastive audio encoders a viable initialization for the audio encoders in next-token prediction models. (4) **ALMs fail to follow instructions.** This is especially true for the complex task of logical reasoning. The next-token prediction ALMs have to be prompted in a specific way, usually matching their training data, to get responses relevant to the user question. If not prompted in a specific way, the ALMs revert to generating text independent of the audio. For example, Pengi's instruction following rate is 61.2% while QwenAudio follows instruction 84.4%, even after matching prompts to training data. This makes it especially challenging to evaluate the ALMs and their responses. We observe that traditional parsing methods are not sufficient to evaluate ALM responses, and hence devise a method to use LLMs to evaluate ALM responses. We setup an ablation study, where we employ human annotators to evaluate ALM as ground truth (details in Appendix). By using LLMs as evaluators we obtain a higher accuracy (96% with Llama3 8B and 99% with Llama3 70B) compared to traditional string parsing or logic methods (70.3%). This LLM evaluator can be further improved along with instruction tuning methods to provide a stronger grounding in audio and instructions.

The highest F1 scores are 51% for the CLE task and 56% for the ACE task, indicating that there is ample room for improving deductive reasoning in contrastive and next-token prediction models.

5.3 Evaluating Audio–Text Representations

The choice of thresholds and prompts used affects ALM performance on the task of entailment. One way to circumvent thresholding and prompting limitations is to evaluate the audio and text representations learned by these models. Therefore, we setup a linear-probe experiment. The audio premise and text hypothesis are encoded by the audio and text encoder, respectively. The audio and text representation are then concatenated and fed to a classifier. In this linear-probe setup, the audio and text encoder are frozen and only the classifier is trained on the target data. We use the CLE dataset, the development set, to train the classifier, the validation to choose the checkpoint, and the test for evaluation.

The linear-probe results are shown in Table 5. The linear-probe leads to an average absolute 30% improvement for Contrastive models while for next-token-prediction we see an absolute improvement of 44%. We can make the observations: (1) The learned audio–text representation can differentiate between possibly true and definitely true, and hence shows primitive reasoning capabilities. The difference between the zero-shot and linear probe performance shows that the current methods of similarity computation and thresholding can be improved. (2) Small parameter count decoder can be compensated by introducing an encoder. This is achieved by using attention throughout audio and instruction (hypothesis), while having autoregressive attention on the suffix. For example, Pengi, which has decoder of 128M, improves performance by having full attention on audio and instruction, while autoregressive attention on output. This aligns with

ALM	Train pairs	ACC↑	P↑	R↑	F1↑	EntACC↑	NeuACC↑	ConACC↑
MS CLAP 22	128k	0.7110	0.7130	0.7110	0.7118	0.6890	0.6775	0.7665
LAION CLAP	2.6M	0.7435	0.7470	0.7435	0.7445	0.7483	0.6957	0.7866
Pengi-enc	3.3M	0.7627	0.7674	0.7627	0.7642	0.7598	0.7100	0.8182
MS CLAP 23	4.6M	**0.8329**	**0.8361**	**0.8329**	**0.8336**	**0.8182**	**0.8440**	**0.8364**

Table 5: Linear-probe performance of Audio Language Models on CLE dataset. Each ALM has an audio encoder and a text encoder to compute embeddings for the audio premise and text hypothesis. The audio embedding and text embedding are concatenated and passed to a linear 3-class classifier.

recent findings in training vision-language models (Beyer et al. 2024). This improves linear-probe performance, but is not effective for zero-shot setup. (3) Despite training the classifier specifically for the audio entailment task, the F1 score remains in the lower 80s. This indicates that the pre-training method could be improved to develop representations capable for logical reasoning.

5.4 Captioning Before Reasoning

Humans employ deductive reasoning by accepting a premise as true, breaking it down into its parts, applying logical principles, and drawing conclusions. Similarly, in audio entailment, models should identify audio events, understand their relationships and order, and infer based on these elements and the hypothesis. This process is similar to creating captions for the audio before engaging in deductive reasoning.

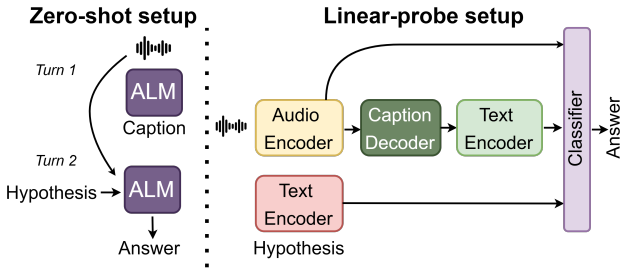

Figure 3: "Caption-before-reason": An intermediate step of audio captioning enhances performance in Audio Entailment tasks. The left figure illustrates a zero-shot setup where ALM is first asked to caption the audio before reasoning with the hypothesis. The right figure depicts a linear probe setup, where a caption and its embedding are generated before being passed to a classifier for prediction.

To evaluate this approach, we conducted two experiments: zero-shot prompting for next-token prediction models and linear probe for contrastive models. We select the best performing model on the CLE dataset, i.e., Qwen-AC, as a representative for next-token prediction models and MS CLAP 2023. For linear probing, we included an explicit audio captioning step using the model's latent embeddings. The generated audio caption was then encoded with a text encoder to produce a sentence-level representation. This encoded hypothesis, along with the caption and base audio representation, was fed into a classifier to make predictions. For zero-shot prompting, we instructed the model to first caption the audio before performing the actual task of audio entailment. We adjust the task prompt to consider both the audio and the generated caption. The setup is illustrated in Figure 3, with results shown in Table 6.

By incorporating an explicit captioning step before making predictions, we observed an absolute improvement in deductive reasoning performance (F1) by 6% for zero-shot prompting and 3% for the linear-probe setup. Using the "caption-before-reason" approach, we observe an increase in accurately predicting contradictions. Previously, the model tended to agree with the hypothesis. However, with explicit captioning, it can better reason and identify misalignments with the audio information. This approach helps the model avoid hallucinating sources based on the hypothesis, and improves grounding in the audio input. Qualitative examples are shown in Appendix. Our prompting approach improves the deductive reasoning performance of ALMs at test-time without requiring training or finetuning.

Model	Method	ACC↑	P↑	R↑	F1↑
Qwen-AC	base	0.5442	0.5604	0.5442	0.4975
Qwen-AC	cap	**0.6083**	**0.5964**	**0.6083**	**0.5601**
CLAP 23	avg	0.7512	0.7529	0.7512	0.7515
CLAP 23	sum	0.7780	0.7812	0.7780	0.7785
CLAP 23	concat	0.8329	0.8361	0.8329	0.8336
CLAP 23	cap	**0.8640**	**0.8671**	**0.8640**	**0.8647**

Table 6: Proposed "caption-before-reason" method for Zero-Shot prompting (top) and linear-probe (bottom).

6 Conclusion

We introduce the Audio Entailment task to evaluate deductive reasoning capabilities of Audio-Language Models (ALMs). We propose two datasets, ACE and CLE, and benchmark state-of-the-art contrastive and next-token prediction ALMs, revealing significant limitations in their logical reasoning abilities. Surprisingly, contrastive models, which learn similarity, performed competitively to next-token prediction models, which learn to produce descriptions. We show limitations of ALMs for following instructions and report quantitative results for the first time in the literature. Finally, we propose "caption-before-reason" to improve zero-shot and linear-probe performance of ALMs by an absolute 6% and 3%, respectively.

References

Anderson, A. R.; Belnap Jr, N. D.; and Dunn, J. M. 2017. *Entailment, Vol. II: The logic of relevance and necessity*, volume 5027. Princeton University Press.

Beyer, L.; Steiner, A.; Pinto, A. S.; Kolesnikov, A.; Wang, X.; Salz, D.; Neumann, M.; Alabdulmohsin, I.; Tschannen, M.; Bugliarello, E.; et al. 2024. PaliGemma: A versatile 3B VLM for transfer. *arXiv preprint arXiv:2407.07726*.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Bukhari, H.; Deshmukh, S.; Dhamyal, H.; Raj, B.; and Singh, R. 2024. SELM: Enhancing Speech Emotion Recognition for Out-of-Domain Scenarios. *arXiv preprint arXiv:2407.15300*.

Chu, Y.; Xu, J.; Zhou, X.; Yang, Q.; Zhang, S.; Yan, Z.; Zhou, C.; and Zhou, J. 2023. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*.

Copi, I. M.; Cohen, C.; and McMahon, K. 2016. *Introduction to logic*. Routledge.

Dagan, I.; Glickman, O.; and Magnini, B. 2005. The pascal recognising textual entailment challenge. In *Machine learning challenges workshop*, 177–190. Springer.

Deshmukh, S.; Elizalde, B.; Singh, R.; and Wang, H. 2023. Pengi: An audio language model for audio tasks. *Advances in Neural Information Processing Systems*, 36: 18090–18108.

Deshmukh, S.; Elizalde, B.; and Wang, H. 2023. Audio Retrieval with WavText5K and CLAP Training. In *Proc. INTERSPEECH 2023*, 2948–2952.

Deshmukh, S.; Singh, R.; and Raj, B. 2024. Domain Adaptation for Contrastive Audio-Language Models. *arXiv preprint arXiv:2402.09585*.

Dhamyal, H.; Elizalde, B.; Deshmukh, S.; Wang, H.; Raj, B.; and Singh, R. 2024. Prompting Audios Using Acoustic Properties for Emotion Representation. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 11936–11940.

Do, V.; Camburu, O.-M.; Akata, Z.; and Lukasiewicz, T. 2020. e-snli-ve: Corrected visual-textual entailment with natural language explanations. *arXiv preprint arXiv:2004.03744*.

Drossos, K.; Lipping, S.; and Virtanen, T. 2020. Clotho: an Audio Captioning Dataset. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Elizalde, B.; Deshmukh, S.; Al Ismail, M.; and Wang, H. 2023. Clap learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

Elizalde, B.; Deshmukh, S.; and Wang, H. 2024. Natural language supervision for general-purpose audio representations. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 336–340. IEEE.

Ghosh, S.; Evuru, C. K. R.; Kumar, S.; Tyagi, U.; Nieto, O.; Jin, Z.; and Manocha, D. 2024a. Visual Description Grounding Reduces Hallucinations and Boosts Reasoning in LVLMs. *arXiv preprint arXiv:2405.15683*.

Ghosh, S.; Kumar, S.; Seth, A.; Evuru, C. K. R.; Tyagi, U.; Sakshi, S.; Nieto, O.; Duraiswami, R.; and Manocha, D. 2024b. GAMA: A Large Audio-Language Model with Advanced Audio Understanding and Complex Reasoning Abilities. *arXiv preprint arXiv:2406.11768*.

Gong, Y.; Liu, A. H.; Luo, H.; Karlinsky, L.; and Glass, J. 2023a. Joint audio and speech understanding. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 1–8. IEEE.

Gong, Y.; Luo, H.; Liu, A. H.; Karlinsky, L.; and Glass, J. 2023b. Listen, think, and understand. *arXiv preprint arXiv:2305.10790*.

Heller, L. M.; Elizalde, B.; Raj, B.; and Deshmukh, S. 2023. Synergy between human and machine approaches to sound/scene recognition and processing: An overview of ICASSP special session. *arXiv preprint arXiv:2302.09719*.

Huang, J.; and Chang, K. C.-C. 2023. Towards Reasoning in Large Language Models: A Survey. In *Findings of the Association for Computational Linguistics: ACL 2023*, 1049–1065.

Kim, C. D.; Kim, B.; Lee, H.; and Kim, G. 2019. AudioCaps: Generating Captions for Audios in The Wild. In *NAACL-HLT*.

Koepke, A. S.; Oncescu, A.-M.; Henriques, J. F.; Akata, Z.; and Albanie, S. 2022. Audio retrieval with natural language queries: A benchmark study. *IEEE Transactions on Multimedia*, 25: 2675–2685.

Kojima, T.; Gu, S. S.; Reid, M.; Matsuo, Y.; and Iwasawa, Y. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35: 22199–22213.

Kong, Z.; Goel, A.; Badlani, R.; Ping, W.; Valle, R.; and Catanzaro, B. 2024. Audio Flamingo: A Novel Audio Language Model with Few-Shot Learning and Dialogue Abilities. *arXiv preprint arXiv:2402.01831*.

Liang, J.; Liu, X.; Liu, H.; Phan, H.; Benetos, E.; Plumbley, M. D.; and Wang, W. 2023. Adapting Language-Audio Models as Few-Shot Audio Learners. arXiv:2305.17719.

Lipping, S.; Sudarsanam, P.; Drossos, K.; and Virtanen, T. 2022. Clotho-AQA: A Crowdsourced Dataset for Audio Question Answering. In *2022 30th European Signal Processing Conference (EUSIPCO)*, 1140–1144.

Liu, F.; Emerson, G. E. T.; and Collier, N. 2023. Visual Spatial Reasoning. *Transactions of the Association for Computational Linguistics*.

MacCartney, B. 2009. *Natural language inference*. Stanford University.

Martín-Morató, I.; and Mesaros, A. 2021. What is the ground truth? reliability of multi-annotator data for audio

tagging. In *2021 29th European Signal Processing Conference (EUSIPCO)*, 76–80. IEEE.

Routley, R.; and Meyer, R. 1973. The semantics of entailment. In *Studies in Logic and the Foundations of Mathematics*, volume 68, 199–243. Elsevier.

Tang, C.; Yu, W.; Sun, G.; Chen, X.; Tan, T.; Li, W.; Lu, L.; MA, Z.; and Zhang, C. 2024. SALMONN: Towards Generic Hearing Abilities for Large Language Models. In *The Twelfth International Conference on Learning Representations*.

Thomas, C.; Zhang, Y.; and Chang, S.-F. 2022. Fine-grained visual entailment. In *European Conference on Computer Vision*, 398–416. Springer.

Wei, J.; Tay, Y.; Bommasani, R.; Raffel, C.; Zoph, B.; Borgeaud, S.; Yogatama, D.; Bosma, M.; Zhou, D.; Metzler, D.; et al. 2022a. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; brian ichter; Xia, F.; Chi, E. H.; Le, Q. V.; and Zhou, D. 2022b. Chain of Thought Prompting Elicits Reasoning in Large Language Models. In Oh, A. H.; Agarwal, A.; Belgrave, D.; and Cho, K., eds., *Advances in Neural Information Processing Systems*.

Wu, Y.; Chen, K.; Zhang, T.; Hui, Y.; Berg-Kirkpatrick, T.; and Dubnov, S. 2023. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

Xie, N.; Lai, F.; Doran, D.; and Kadav, A. 2019. Visual entailment: A novel task for fine-grained image understanding. *arXiv preprint arXiv:1901.06706*.